Embedding Regression:
Models for Context-Specific Description and Inference*

Pedro L. Rodriguez† Arthur Spirling‡ Brandon M. Stewart§

11534 words

Abstract

Social scientists commonly seek to make statements about how word use varies over circumstances—including time, partisan identity, or some other document-level covariate. For example, researchers might wish to know how Republicans and Democrats diverge in their understanding of the term "immigration." Building on the success of pretrained language models, we introduce the *à la Carte on Text* (`conText`) embedding regression model for this purpose. This fast and simple method produces valid vector representations of how words are used—and thus what words "mean"—in different contexts. We show that it outperforms slower, more complicated alternatives, and works well even with very few documents. The model also allows for hypothesis testing and statements about statistical significance. We demonstrate that it can be used for a broad range of important tasks, including understanding US polarization, historical legislative development, and sentiment detection. We provide open-source software for fitting the model.

---

*First draft: July 2020. This draft: June 1, 2022. We thank audience members at the Midwest Political Science Association Annual Meeting (2021), the Political Methodology society meeting, the American Political Science Association Annual Meeting (2021), Princeton University, and the University of Wisconsin (Madison). We are grateful to Clark Bernier, Saloni Bhogale, Max Goplerud, Justin Grimmer, Alex Kindel, Hauke Licht, John Londregan, Walter Mebane and Molly Roberts for comments. We also thank the editor and four excellent anonymous reviewers for their careful engagement with our work.

†Postdoctoral Fellow, Data Science Institute (joint with Political Science), Vanderbilt University and International Faculty at Instituto de Estudios Superiores de Administración (`pedro.rodriguez@Vanderbilt.Edu`)

‡Professor of Politics and Data Science, New York University (`arthur.spirling@nyu.edu`)

§Assistant Professor, Sociology and Office of Population Research, Princeton University (`bms4@princeton.edu`)

1 Introduction

All human communication requires common understandings of meaning. This is nowhere more true than political and social life, where the success of an appeal—rhetorical or otherwise—relies on an audience perceiving a message in the particular way that the speaker seeks to deliver it. Scholars have therefore spent much effort exploring the meanings of terms, how those meanings are manipulated, and how they change over time and space. Historically, this work has been qualitative (e.g. Austin, 1962; Skinner, 1969; Geertz, 1973). But in recent times, quantitative analysts have turned to modeling and measuring "context" directly from natural language (e.g. Hopkins, 2018; Aslett et al., 2022; Park, Greene and Colaresi, 2020).

A promising avenue for such investigations has been the use of "word embeddings"—a family of techniques that conceive of meaning as emerging from the distribution of words that surround a term in text (e.g. Mikolov et al., 2013). By representing each word as a vector of real numbers, and examining the relationships between vectors for the vocabulary of a corpus, scholars have uncovered new facts about language and the people that produce it (e.g. Caliskan, Bryson and Narayanan, 2017). This is also true in the study of politics, society and culture (Garg et al., 2018; Kozlowski, Taddy and Evans, 2019; Rodman, 2020; Rheault and Cochrane, 2020; Wu et al., 2019).

While borrowing existing techniques has certainly produced insights, for social scientists two problems remain. First, traditional approaches generally require a lot of data to produce high quality representations—that is, to produce embeddings that make sense and connote meaning of terms correctly. The issue is less that our typical corpora are small—though they are compared to those on the web-scale collections often used in computer science—and more that terms for which we would like to estimate contexts are subject-specific and thus typically quite *rare*. As an example, there are fewer than twenty parliamentary mentions of the "special relationship" between the US and the UK in some years of the 1980s—despite this arguably being the high watermark of elite closeness between the two countries. The second problem is one of inference. While representations themselves are helpful, social scientists want to make statements about the statistical properties and relationships between embeddings. That is, they want to speak meaningfully of whether language is used differently across subcorpora and whether those apparent differences are larger than we would expect by chance. Neither of these problems are well-addressed by current techniques. While there have been efforts to address inference in embeddings (see, e.g, Kulkarni et al., 2015; Lauretig, 2019), they are typically data intensive and computationally intensive.

We tackle these two problems together in what follows. We provide both a statistical framework for making statements about covariate effects on embeddings, and one that performs particularly well in cases of rare words or small corpora. Specifically, we innovate on Khodak et al. (2018) which introduced *à la carte embeddings* (ALC). In a nutshell, the method takes embeddings which have been pre-trained on large corpora (e.g. `word2vec` or `GloVe` embeddings readily available online), combined with a small sample of example uses for a focal word, and then induces a new context-specific embedding for the focal word. This requires only a simple linear transformation of the averaged embeddings for words within the context of the focal word.

We place ALC in a *regression* setting that allows for fast solutions to queries like "do authors with these covariate values use these terms in a different way than authors with different covariate values? If yes, how do they differ?" We provide three proofs of concept. First, we demonstrate the strength of our approach by comparing its performance to the "industry standard" as laid out by Rodman (2020) in a study of a New York Times corpus, where slow changes over long periods are the norm. Second, we show that our approach can estimate an approximate embedding even with only a single context. In particular, we demonstrate that we can separate individual instances of `Trump` and `trump`. Third, we show that our method can also identify drastic switches in meaning over short time periods—specifically in our case, for the term `Trump` before and after the 2016 election.

We study three substantive cases to show how the technique may be put to work. First, we explore partisan differences in Congressional speech—a topic of long-standing interest in political science (see, e.g., Monroe, Colaresi and Quinn, 2008). We show that `immigration` is, perhaps unsurprisingly, one of the most differently expressed terms for contemporary Democrats and Republicans. Our second substantive case is historical: we compare across polities (and corpora) to show how elites in the UK and US expressed `empire` in the post-war period, how that usage diverged, and when. Our third case shows how our approach can be used to measure sentiment. We build on earlier work (e.g. Slapin et al., 2018; Osnabrügge, Hobolt and Rodon, 2021) for the UK House of Commons, yielding novel insights about the relationship between the UK Prime Minister and his backbenchers on the European Union. We also provide advice to practitioners on how to use the technique based on extensive experiments reported in the Supporting Information (SI).

These innovations allow for social scientists to go beyond general meanings of words to capture situation-specific usage. This is possible without substantial computation and, in contrast to other approaches, requires only the text immediately around the word of interest.

We proceed as follows: in Section 2 we provide some context for what social scientists mean by 'context' and link this to the distribution of words around a focal term. We then introduce the ALC algorithm, and provide three proofs of concept. Subsequently, we extend ALC to a regression framework, and then present results from three substantive use-cases. We give practical guidance on use and limitations before concluding.

2 Context in Context

> …they are casting their problems on society and who is society? There is no such thing!
>
> —Margaret Thatcher, interview with *Woman's Own* (1987).

Paraphrased as "there is no such thing as society", Thatcher's quote has produced lively debate in the study and practice of UK politics. Critics—especially from the left—argued that this was primarily an endorsement of individual selfishness and greed. But more sympathetic accounts have argued that the quote must be seen in its full *context* to be understood. The implication is that reading the line in its original surroundings changes the meaning: rather than embracing egotism, it emphasizes the importance of citizens' obligations to each other above and beyond what the state requires.

Beyond this specific example, the measurement and modeling of "context" is obviously a general problem. In a basic sense, context is vital: we literally cannot understand what is meant by a speaker or author without it. This is partly due to polysemy—the word "society" might mean many different things. But the issue is broader than this and is at the core of human communication. Unsurprisingly then, the study of context has been a long-standing endeavor in social science. Its centrality has been emphasized in the history of ideas (Skinner, 1969) through the lens of "speech acts" (Austin, 1962); describing cultural practices via "thick description" (Geertz, 1973); understanding "political culture" (Verba and Almond, 1963); and the psychology of decision making (Tversky and Kahneman, 1981).

2.1 Approaches to Studying Context

For the goal of describing context in observational data, social science has turned to text approaches—with topic models being popular (see Grimmer, 2010; Quinn et al., 2010; Roberts, Stewart and Airoldi, 2016). Topic models provide a way to understand the allocation of attention across groupings of words.

While such models have a built-in notion of polysemy (a single word can be allocated to different topics), they are rarely used as a mechanism for studying how individual words are used to convey different ideas (Grimmer and Stewart, 2013). And though topic approaches do exist that allow for systematic variation in the use of a word across topics by different pieces of observed metadata (Roberts, Stewart and Airoldi, 2016), they are computationally intensive (especially relative to the approaches we present below). The common unit of analysis for topic models is the document. This has implications for the way that these models capture the logic of the "Distributional Hypothesis"—the idea that, in the sense of Firth (1957, 11), "You shall know a word by the company it keeps"—i.e. that one can understand a particular version of the "meaning" of a term from the way it co-occurs with other terms. Specifically, in the case of topic models, the entire document is the context. From this we learn the relationships (the themes) between words and the documents in which they appear.

But in the questions we discuss here, the interest is in the contextual use of a *specific word*. To study this, social scientists have turned to word embeddings (e.g. Rheault and Cochrane, 2020; Rodman, 2020). For example, Caliskan, Bryson and Narayanan (2017) and Garg et al. (2018) have explored relationships between words captured by embeddings to describe problematic gender and ethnic stereotypes in society at large. These approaches predict a focal word as a function of the other words that appear within a small window of that focal word in the corpus (or the reverse, predict the neighboring words from the focal word). In so doing, they capture the insight of the Distributional Hypothesis in a very literal way: the "context" of a term are the tokens that appear near it in text, on average. In practice, this is all operationalized via a matrix of co-occurrences of words that respect the relevant window size. In the limit, where we imagine the relevant window is the entire document, one can produce a topic model from the co-occurrence matrix directly. Thus as the context window in the embedding model approaches the length of the document, the embeddings will increasingly look like the word representations in a topic model.

Whether, and in what way, embedding models based on the Distributional Hypothesis capture "meaning" is more controversial. Here we take a narrow, "structuralist" (in the sense of Harris, 1954) view. For this paper, meaning is in terms of *description* and is empirical. That is, it arises from word co-occurrences in the data, alone: we will not construct, nor assume, a given theoretical model of language or cognition. And, in contrast to other scholars (e.g. Miller and Charles, 1991), we will make no claims that the distributions *per se* have causal effects on human understandings of terms. Thus, when we speak of the meaning of a focal word being different across groups, we are talking in a thin sense about the distribution of other words within a fixed window size of that focal word being different. Though we will offer guidance, substantive interpretation of these differences for a given purpose is ultimately up to the researcher. That is, as always with such text measurement strategies, subject-expert validation is important.

For a variety of use-cases, social scientists want to make systematic inferences about embeddings—which requires statements about uncertainty. Suppose we wish to compare the context of "society" as conveyed by British Prime Ministers with that of US Presidents. Do they differ in a statistically significant way? To judge this, we need some notion of a null hypothesis, some understanding of the variance of our estimates, and a test statistic. While there have been efforts to compare embeddings across groups (Rudolph et al., 2017), and to give frameworks for such conditional relationships (Han et al., 2018), these are non-trivial to implement. Perhaps more problematically for most social science cases, they rely on underlying embedding models that struggle to produce "good" representations—that make sense, and correctly capture how that word is actually used—when we have few instances of a term of interest. This matters because we are typically far short of the word numbers which standard models require for optimal performance and terms (like "society") may be used in ways that are idiosyncratic to a particular document or author.

In the next section, we will explain how we build on earlier insights from ALC embeddings (Khodak et al., 2018) to solve these problems in a fast, simple, and sample-efficient "regression" framework. Before doing so, we note three substantive use cases that both motivate the methodological work we do, and show its power as a tool for social scientists. The exercise in all cases is linguistic *discovery* insofar as our priors are not especially sharp, and the primary value is in stimulating more productive engagement with the text. Nonetheless, in using the specific approach we outline in this paper, we will be able to make inferences with attendant statements about uncertainty. In that sense, our examples are intended to be illuminating for other scholars comparing corpora or comparing authors within a corpus.

Use-case I: Partisan Differences in word usage. A common problem in Americanist political science is to estimate partisan differences in the usage of a given term. Put literally: do Republicans and Democrats mean something different when they use otherwise identical words like `immigration` and `marriage`? While there have been efforts to understand differential word *rate of use* within topics pertaining to these terms (e.g. Monroe, Colaresi and Quinn, 2008), there has been relatively little work on whether the *same* words appear in different contexts. Below, we use the *Congressional Record* (Sessions 111–114) as our corpus for this study (Gentzkow, Shapiro and Taddy, 2018). This requires that we compare embeddings as a function of party (and other covariates).

Use-case II: Changing UK-US Understandings of '`Empire`'. The United Kingdom's relative decline as a Great Power in the post-war period has been well-documented (e.g. Hennessy, 1992). One way that we might investigate the timing of US dominance (over the UK, at least) is to study the changing understanding of the term "`Empire`" in both places. That is, beyond any attitudinal shift, did American and British policy-makers alter the way they used empire as the century wore on? If they did, when did this occur? And did the elites of these countries converge or diverge in terms of their associations of the term? To answer these questions, we will statistically compare the embedding for the term "`Empire`" for the UK House of Commons (via *Hansard*) versus the US Congress (via the *Congressional Record* from 1935–2010.

Use-case III: Brexit Sentiment from the Backbenches The UK's decision to leave the European Union (EU) following the 2016 referendum was momentous (Ford and Goodwin, 2017). While the vote itself was up to citizens, the build-up to the plebescite was a matter for elites; specifically, it was a consequence of the internal machinations of the parliamentary Conservative Party that forced the hand of their leader, Prime Minster David Cameron (Hobolt, 2016). A natural question concerns the attitudes of that party in the House of Commons towards the EU, both over time and relative to other issue areas (such as education and health policy). To assess that, we will use an embedding approach to sentiment estimation for *single instances* of terms, that builds on recent work on emotion in parliament (Osnabrügge, Hobolt and Rodon, 2021). This will also allow us to contribute to the literature on Member of Parliament (MP) position-taking via speech (see, e.g. Slapin et al., 2018).

3 Using ALC Embeddings To Measure Meaning

Our methodological goal is a regression framework for embeddings. By "regression" we mean two related ideas. Narrowly, we mean that we want to be able to approximate a *conditional expectation function*, typically written $\mathbb{E}[Y|X]$ where, as usual, Y is our outcome, X is a particular covariate, and $\mathbb{E}$ is the expectations operator. We want to make statements about how embeddings (our Y) differ as covariates (our X) change. More broadly, we use "regression" to mean machinery for testing hypotheses about whether the groups actually differ in a systematic way. And by extension, we want that machinery to provide tools for making downstream comments about *how* those embeddings differ. In all cases, this will require three related operations:

1. an efficient and transparent way to embed words, such that we can produce high quality representations even when a given word is rare.

2. given (1), a demonstration that in real problems, a *single* instance of a word's use is enough to produce a good embedding. This allows us to set up the hypothesis-testing problem as a multivariate regression, and is the subject of Section 4.1.

3. given (1) and (2), a method for making claims about the statistical significance of differences in embeddings, based on covariate profiles. We tackle that in Section 4.3.

Ideally, our framework will deliver good representations of meaning even in cases where we have very few incidences of the words in question. ALC embeddings (Khodak et al., 2018) promise exactly this. We now give some background and intuition on that technique. We then replicate Rodman (2020)—a recent study introducing time-dependent word embeddings for political science—to demonstrate ALC's efficiency and quality.

3.1 Word Embeddings Measure Meaning Through Word Co-Occurence

Word embeddings techniques give every word a *distributed representation*—that is, a vector. The length or dimension (D) of this vector is—by convention—between 100 and 500. When the inner product between two different words (two different vectors) is high, we infer that they are likely to co-occur in similar contexts. The Distributional Hypothesis then allows us to infer that those two words are similar in *meaning*. While such techniques are not new conceptually (e.g. Hinton, 1986), methodological advances in the last decade (Mikolov et al., 2013; Pennington, Socher and Manning, 2014) allow them to be estimated much more quickly. More substantively, word embeddings have been shown to be useful, both as inputs to supervised learning problems and for understanding language directly. For example, embedding representations can be used to solve analogy reasoning tasks, implying the vectors do indeed capture relational meaning between words (e.g. Arora et al., 2018).

Understanding exactly *why* word embeddings work is non-trivial. In any case, there is now a large literature proposing variants of the original techniques (e.g. Faruqui et al., 2015; Lauretig, 2019). A few of these are geared specifically to social science applications where the general interest is in measuring changes in meanings, especially via "nearest neighbors" of specific words.

While the learned embeddings provide a rough sense of what a word means, it is difficult to use them to answer questions of the sort we posed above. Consider our interest in how Republicans and Democrats use the same word (e.g. `immigration`) differently. If we train a set of word embeddings on the entire *Congressional Record* we only have a single meaning of the word. We could instead train a separate set of embeddings—one for Republicans and one for Democrats—and then realign them. This is an extra computational step, and may not be feasible in other use cases where the vocabularies do not have much overlap. We now discuss a way to proceed that is considerably easier.

3.2 A Random Walk Theoretical Framework and ALC Embeddings

The core of our approach are ALC embeddings. The theory behind that approach is given by Arora et al. (2016) and Arora et al. (2018). Those papers conceive of documents being a 'random walk' in a discourse space, where words are more likely to follow other words if they are closer to them in an embedding space. Crucially for ALC, Arora et al. (2018) also proves that under this model, a particular relationship will follow for the embedding of a word and the embeddings of the words that appear in the contexts *around it.*

To fix ideas, consider the following toy example. Our corpus is the memoirs of a politician, and we observe two entries, both mentioning the word 'bill':

1. *The debate lasted hours, but finally we* [voted on the $\boxed{\text{bill}}$ and it passed] *with a large majority.*

2. *At the restaurant we ran up* [a huge wine $\boxed{\text{bill}}$ to be paid] *by our host.*

As one can gather from the context—here, the three words either side of the instance of 'bill' in square brackets—the politician is using the term in two different (but grammatically correct) ways.

The key result from Arora et al. (2018) shows the following: if the random walk model holds, the researcher can obtain an embedding for word w (e.g. 'bill') by taking the average of the embeddings of the words around w ($\mathbf{u}_w$) and multiplying them by a particular square matrix $\mathbf{A}$. That $\mathbf{A}$ serves to downweight the contributions of very common (but uninformative) words when averaging. Put otherwise, if we can take averages of some vectors of words that surround w (based on some pre-existing set of embeddings) and if we can find a way to obtain $\mathbf{A}$ (which we will see is also straightforward), we can provide new embeddings for even very rare words. And we can do this almost instantaneously.

Returning to our toy example, consider the first, legislative, use of 'bill' and the words around it. Suppose we have embedding vectors for those words from some other larger corpus, like Wikipedia. To keep things compact, we will suppose those embeddings are all of three dimensions (such that $D = 3$), and take the following values:

$$\underbrace{\begin{bmatrix} -1.22 \\ 1.33 \\ 0.53 \end{bmatrix}}_{\text{voted}} \underbrace{\begin{bmatrix} 1.83 \\ 0.56 \\ -0.81 \end{bmatrix}}_{\text{on}} \underbrace{\begin{bmatrix} -0.06 \\ -0.73 \\ 0.82 \end{bmatrix}}_{\text{the}} \quad \text{bill} \quad \underbrace{\begin{bmatrix} 1.81 \\ 1.86 \\ 1.57 \end{bmatrix}}_{\text{and}} \underbrace{\begin{bmatrix} -1.50 \\ -1.65 \\ 0.48 \end{bmatrix}}_{\text{it}} \underbrace{\begin{bmatrix} -0.12 \\ 1.63 \\ -0.17 \end{bmatrix}}_{\text{passed}}$$

Obtaining $\mathbf{u}_w$ for 'bill' simply requires averaging these vectors and thus

$$\mathbf{u}_{\text{bill}_1} = \begin{bmatrix} 0.12 \\ 0.50 \\ 0.40 \end{bmatrix},$$

with the subscript denoting the first use case. We can do the same for the second case—the restaurant sense of 'bill'—from the vectors of `a`, `huge`, `wine`, `to`, `be` and `paid`. We obtain

$$\mathbf{u}_{\text{bill}_2} = \begin{bmatrix} 0.35 \\ -0.38 \\ -0.24 \end{bmatrix},$$

which differs from the average for the first meaning. A reasonable instinct is that these two vectors should be enough to give us an embedding for 'bill' in the two senses. Unfortunately, they will not—this is shown empirically in Khodak et al. (2018) and in our `Trump`/`trump` example below. As implied above, the intuition is that simply averaging embeddings over-exaggerates common components associated with frequent (e.g. "stop") words. So we will need the **A** matrix too: it downweights these directions so they don't overwhelm the induced embedding.

Khodak et al. (2018) show how to put this logic into practice. The idea is that a large corpus (generally the corpus the embeddings were originally trained on, such as Wikipedia) can be used to estimate the transformation matrix $\mathbf{A}$. This is a one time cost after which each new word embedding can be computed *à la carte* (hence the name), rather than needing to retrain an entire corpus just to get the embedding for a single word. As a practical matter, the estimator for $\mathbf{A}$ can be learned efficiently with a lightly modified linear regression model which reweights the words by a non-decreasing function $\alpha(\cdot)$ of the total instances of each word (n_w) in the corpus. This reweighting addresses the fact that words which appear more frequently have embeddings which are measured with greater certainty. Thus we learn the transformation matrix as,

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \sum_{w=1}^{W} \alpha(n_w) \|\mathbf{v}_w - \mathbf{A}\mathbf{u}_w\|_2^2 \tag{1}$$

The natural log is a simple choice for $\alpha(\cdot)$, and works well. Given $\hat{\mathbf{A}}$, we can introduce new embeddings for any word by averaging the existing embeddings for all words in its context to create $\mathbf{u}_w$ and then applying the transformation such that $\hat{\mathbf{v}}_w = \hat{\mathbf{A}}\mathbf{u}_w$. The transformation matrix is not particularly hard to learn (it is a linear regression problem) and each subsequent induced word embedding is a single matrix multiply.

Returning to our toy example, suppose that we estimate $\hat{\mathbf{A}}$ from a large corpus like *Hansard* or the *Congressional Record* or wherever we obtained the embeddings for the words that surround 'bill.' Suppose that we estimate

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.81 & 3.96 & 2.86 \\ 2.02 & 4.81 & 1.93 \\ 3.14 & 3.81 & 1.13 \end{bmatrix}.$$

Taking inner products, we have

$$\mathbf{v}_{\text{bill}_1} = \hat{\mathbf{A}} \cdot \mathbf{u}_{\text{bill}_1} = \begin{bmatrix} 3.22 \\ 3.42 \\ 2.73 \end{bmatrix} \quad \text{and} \quad \mathbf{v}_{\text{bill}_2} = \hat{\mathbf{A}} \cdot \mathbf{u}_{\text{bill}_2} = \begin{bmatrix} -1.91 \\ -1.58 \\ -0.62 \end{bmatrix}.$$

These two transformed embeddings vectors are more different than they were—a result of downweighting the commonly appearing words around them—but that is not the point *per se*. Rather, we expect them to be informative about the word sense by, for example, comparing them to other (pre-estimated) embeddings in terms of distance. Thus we might find that the nearest neighbors of $\mathbf{v}_{\text{bill}_1}$ are

$$\text{legislation} = \begin{bmatrix} 3.11 \\ 2.52 \\ 3.38 \end{bmatrix} \quad \text{and} \quad \text{amendment} = \begin{bmatrix} 2.15 \\ 2.47 \\ 3.42 \end{bmatrix}$$

while the nearest neighbors of $\mathbf{v}_{\text{bill}_2}$ are

$$\text{dollars} = \begin{bmatrix} -1.92 \\ -1.54 \\ -0.60 \end{bmatrix} \quad \text{and} \quad \text{cost} = \begin{bmatrix} -1.95 \\ -1.61 \\ -0.63 \end{bmatrix}.$$

This makes sense, given how we would typically read the politician's lines above. The key here is that the ALC method allowed us to infer the meaning of words that occurred rarely in a small corpus (the memoirs) without having to build embeddings for those rare words in that small corpus: we could 'borrow' and transform the embeddings from another source. Well beyond this toy example, Khodak et al. (2018) finds empirically that the learned $\hat{\mathbf{A}}$ in a large corpus recovers the original word vectors with high accuracy (greater than .9 cosine similarity). They also demonstrate that this strategy achieves state-of-the-art and near state-of-the-art performance on a wide variety of natural language processing tasks (e.g. learning the embedding of a word using only its definition, learning meaningful n-grams, classification tasks etc.) at a fraction of the computational cost of the alternatives.

The ALC framework has three major advantages for our setting: transparency, computational ease, and efficiency. First, compared to many other embedding strategies for calculating conditional embeddings (e.g., words over time) the information used in ALC is transparent. The embeddings are derived directly from the additive information of the words in the context window around the focal word, there is no additional smoothing or complex interactions across different words. Furthermore, the embedding space itself does not change, it remains fixed to the space defined by the pre-trained embeddings. Second, this same transparency leads to computational ease. The transformation matrix $\mathbf{A}$ only has to be estimated once and then each subsequent induction of a new word is a single matrix multiply and thus effectively instantaneous. Later we will be able to exploit this speed to allow bootstrapping and permutation procedures that would be unthinkable if there was an expensive model fitting procedure for each word. Finally, ALC is efficient in the use of information. Once the transformation matrix is estimated, it is only necessary that $\mathbf{u}_w$ converges—in other words, we only need to estimate a D-dimensional mean from a set of samples. In the case of a 6-word symmetric context window there are twelve words total within the context window; thus, for each instance of the focal word we have a sample of size 12 from which to estimate the mean.

While Khodak et al. (2018) focused on using the ALC framework to induce embeddings for rare words and phrases, we will apply this technique to embed words used in different partitions of a single corpus or to compare across corpora. This allows us to capture differences in embeddings over time or by speaker, even when we have only a few instances within each sample. Importantly, unlike other methods, we don't need an entirely new corpus to learn embeddings for select focal words, we can select particular words and calculate (only) their embeddings using only the contexts around those particular words.[1] We now demonstrate this power of ALC by replicating Rodman (2020).[2]

3.3 Proof of Concept for ALC in Small Political Science Corpora: Reanalyzing Rodman (2019)

The task in Rodman (2020) is to understand changes in the meaning of `equality` over the period 1855–2016 in a corpus consisting of the headlines and other summaries of news articles. As a gold standard, a subset of the articles is hand-coded into fifteen topic word categories—of which five are ultimately used in the analysis—and the remaining articles are coded using a supervised topic model with the hand-coded data as input. Four embeddings techniques are used to approximate trends in coverage of those categories, via the (cosine) distance between the embedding for the word `equality` and the embeddings for the category labels. This is challenging, because the corpus is small—the first 25 year slice of data has only 80 documents—and in almost 30% of the word-era combinations there are fewer than 10 observations.[3]

Rodman (2020) tests four different methods by comparing results to the gold standard; ultimately, the chronologically trained model (Kim et al., 2014) is the best performer. In each era (of 25 years), the model is fit several times on a bootstrap resampled collection

[1]For context, there are many approaches in computer science including anchoring words (Yin, Sachidananda and Prabhakar, 2018) and vector space alignment (Hamilton, Leskovec and Jurafsky, 2016).

[2]Many papers in computer science have studied semantic change (see Kutuzov et al., 2018, for a survey).

[3]We provide more information on the sample constraints in Supporting Information, Part A.

of documents and then averaged over the resulting solutions (Antoniak and Mimno, 2018). Importantly, the model in period t is initialized with period $t-1$ embeddings, while the first period is initialized with vectors trained on the full corpus. Even for a relatively small corpus this process is computationally expensive, and our replication took about five hours of compute time on an 8-core machine.

The ALC approach to the problem is simple. For each period we use ALC to induce a period-specific embedding for `equality` as well as each of the five category words: `gender`, `treaty`, `german`, `race` and `african_american`. We use `GloVe` pre-trained embeddings and the corresponding transformation matrix estimated by Khodak et al. (2018)—in other words, we make use of no corpus-specific information in the initial embeddings and require as inputs *only the context window around each category word*. Following Rodman, we compute the cosine similarity between `equality` and each of the five category words, for each period. We then standardize (make into z-scores) those similarities. The entire process is transparent and takes only a few milliseconds (the embeddings themselves involve six matrix multiplies).

How does ALC do? Figure 1 is the equivalent of Figure 3 in Rodman (2020). It displays the normalized cosine similarities for the chronological model (CHR, taken from Rodman (2020)) and ALC, along with the gold standard (GS). We observe that ALC tracks approximately as well as Rodman's chronological model on its own terms. Where ALC clearly does better is on each model's nearest neighbors (Tables 1 and 2): it produces more semantically interpretable and conceptually precise nearest neighbors than the chronological model. This is partly a result of the ALC model being able to produce nearest neighbors beyond those in the original corpus, borrowing from semantic information stored in the pre-trained embeddings.

We emphasize that in the 1855 corpus, four of the five category words (all except `african_american`) are estimated using *five or fewer instances*. While the chronological model is sharing information across time periods, ALC is treating each slice separately, meaning that our analysis could be conducted effectively with even fewer time periods.[4]

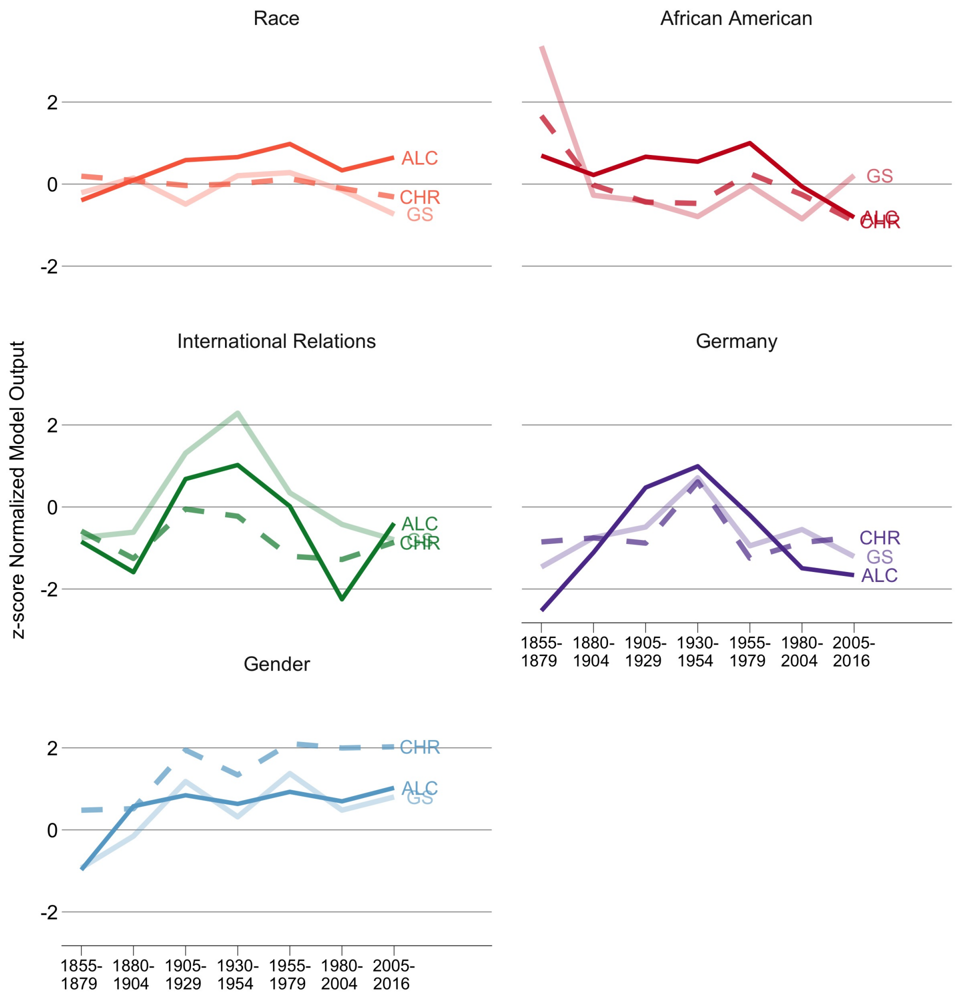


Figure 1: Replication of Figure 3 in Rodman (2020) adding ALC results. ALC = ALC model; CHR = chronological model and GS = gold standard.

[4]An advantage of our embedding regression framework is that we can also share information across time periods using a regression model without sacrificing any of the simplicity or speed of ALC.

african_american		gender		treaty		german		race		equality	
CHR	ALC	CHR	ALC	CHR	ALC	CHR	ALC	CHR	ALC	CHR	ALC
equality	suffrage	will	legislatures	britain	equality	reich	visit	enfranchisement	enfranchisement	of	enacment
the	emancipation	performing	missourians	extradition	toleration	berlin	france	marriage	equality	the	abolition
and	fairness	give	suffrage	interpolation	speech	arms	eugenia	newmarket	interrelation	and	enacting
of	guaranteeing	blackwell	disestablish	minister	championing	hitler	bilateral	louise	expounder	in	effecting
whites	slavery	american	constitutions	rouher	extradition	von	relations	need	abrogation	to	abolishment

Table 1: Nearest neighbors for the 1855 corpus.

african_american		gender		treaty		german		race		equality	
CHR	ALC	CHR	ALC	CHR	ALC	CHR	ALC	CHR	ALC	CHR	ALC
crandall's	nonwhites	equality	equality	narrow	equality	maintains	universities	universe	equality	the	gender
costs	asians	the	inequalities	designed	affirms	hinge	colleges	1950s	segregation	for	gays
unraveling	cubans	for	inequity	missed	reaffirms	holstein's	campuses	warriors	inequalities	of	lesbians
treats	suburbanites	of	inequality	assure	affirming	equality's	striving	posits	discrimination	and	transgender
congresswoman	championing	and	lesbians	trade	upholds	kiel	decades	purdy'	affirmative	to	lgbt

Table 2: Nearest neighbors for the 2005 corpus.

Collectively, these results suggest that ALC is competitive with the current state of the art within the kind of small corpora that arise in social science settings. We now turn to providing a hypothesis testing framework that will allow us to answer the types of questions we introduced above.

4 Testing Hypotheses about Embeddings

Ultimately we want to speak of the way that embeddings differ systematically across levels of covariates. To do this, we will set up a regression-like framework, where each 'observation' is the embedding of a single word. ALC will assist us, but first we show that it can learn meaningful embeddings from *one* example use.

4.1 ALC Can Distinguish Word Meanings From One Example Use

Above we explained that ALC averaged pre-trained embeddings and then applied a linear transformation. This new embedding vector has, say, 300 dimensions, and we might reasonably be concerned that it is too noisy to be useful. To evaluate this, we need a ground truth. So we study a recent *New York Times* corpus; based on lead paragraphs, we show that we can reliably distinguish `Trump` the person (2017–2020) from other sense of `trump` as a verb or noun (1990–2020).

For each sense of the word (based on capitalization) we take a random sample of 100 realizations from our New York Times corpus and embed them using ALC. We apply k-means clustering with two clusters to the set of embedded instances and evaluate whether the clusters partition the two senses. If ALC works, we should obtain two separate clouds of points that are internally consistent (in terms of the senses of the term). This is approximately what we see. Figure 2 provides a visualization of the 300-dimensional space projected to two dimensions with PCA and identifying the two clusters by their dominant word sense. We explicitly mark misclassifications with an x.

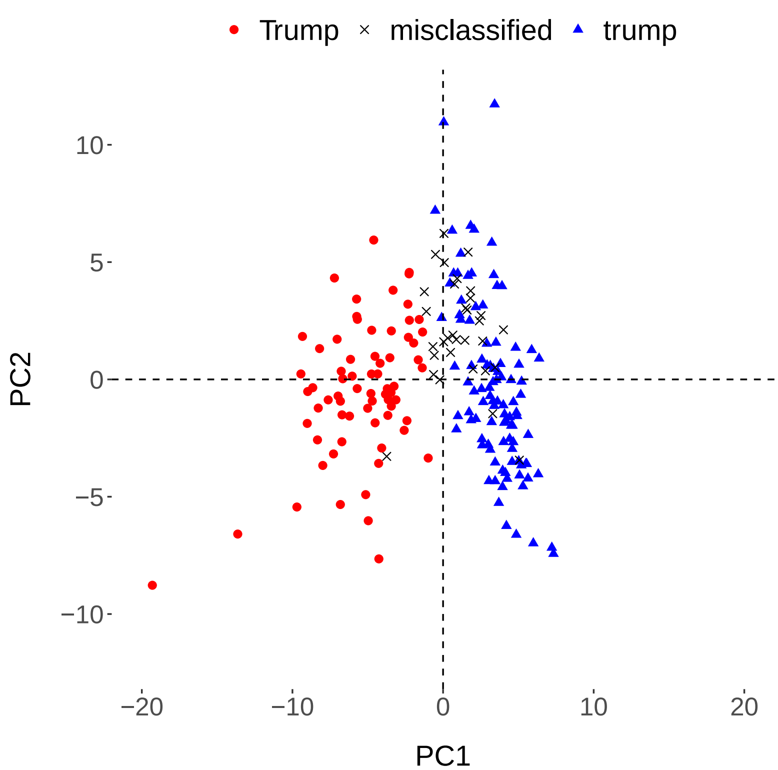


Figure 2: Each observation represents a single realization of a target word, either of `trump` or `Trump`. Misclassified instances refer to instances of either target word that were assigned the majority cluster of the opposite target word.

To provide a quantitative measure of performance we compute the average cluster homogeneity: the degree to which each cluster contains only members of a given class. This value ranges between 0—both clusters have equal numbers of both context types—and 1—each cluster consists entirely of a single context type. By way of comparison, we do the same exercise using other popular methods of computing word vectors for each target realization including: latent semantic analysis (LSA), simple averaging of the corresponding pre-trained embeddings (ALC without transformation by **A**) and RoBERTa contextual embeddings (Liu et al., 2019).[5][6] To quantify uncertainty in our metric, we use block bootstrapping—resampling individual instances of the focal word.[7] Figure 3 summarizes our results.

LSA does not fare well in this task. ALC, on the other hand, performs close to on par with transformer-based RoBERTa embeddings.[8] Simple averaging of embeddings also performs surprisingly well, coming out on top in this comparison. Does this mean the linear transformation that distinguishes ALC from simple averaging is redundant? To evaluate this, we look at nearest neighbors using both methods. Table 3 displays these results. We observe

---

[5]For LSA we use two dimensions and tf-idf weighting. We found these settings to produce the best results.

[6]RoBERTa is a substantially more complicated embedding method which produces contextually-specific embeddings and uses word order information.

[7]Note here that we are treating the **A** matrix as fixed and thus we are not incorporating uncertainty in those estimates. In experiments (see Supporting Information, Part F) we found this uncertainty to be minor and a second-order concern for our applications.

[8]This may be a result of RoBERTa being optimized for sentence embeddings more than embeddings for an individual word. Nonetheless, it is surprising given that transformer-based models lead almost every natural language process benchmark task. Even at comparable performance though there would be reason not to use RoBERTa models simply based on computational cost and comparative complexity.

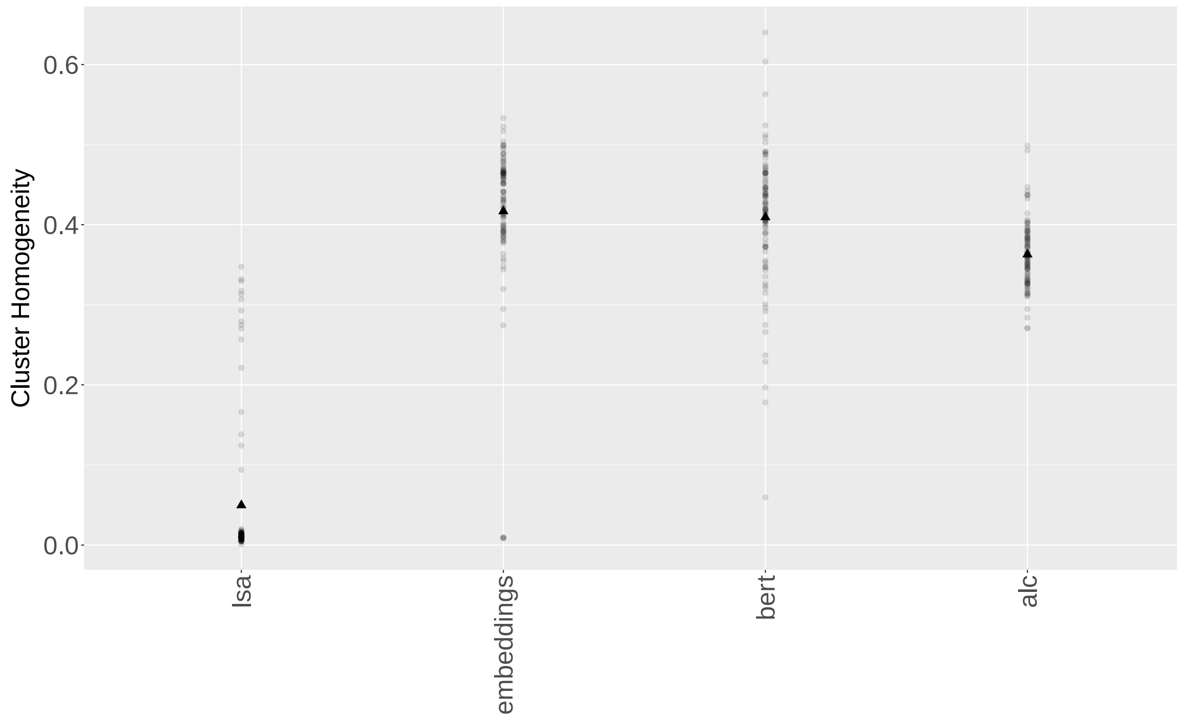


Figure 3: Cluster homogeneity (in terms of `Trump` vs. `trump`) of k-means with 2 clusters of individual term instances embedded using different methods.

that simple averaging of embeddings produces mainly stopwords as nearest neighbors. ALC, on the other hand, outputs nearest neighbors aligned with the meaning of each term, `Trump` is associated with president Trump while `trump` is largely associated with its two related other meanings: a suit in trick-taking games and defeating someone. This serves to highlight the importance of the linear transformation $\mathbf{A}$ in the ALC method.

While this example is a relatively straightforward case of polysemy, we also know that the meaning of `Trump`, the surname, underwent a significant transformation once Donald J. Trump was elected president of the United States in November 2016. This is a substantially harder case since the person being referred to is still the same, even though the contexts it is employed in—and thus in the sense of the distributional hypothesis, the meaning—has shifted. But as we show in Supporting Information B, ALC has no problem with this case either, returning excellent cluster homogeneity and nearest neighbors.

The good news for the `Trump` examples is that ALC can produce reasonable embeddings

Trump		trump	
Embeddings	**ALC**	**Embeddings**	**ALC**
but	president	but	declarer
that	assailed	only	spades
even	clinton	even	colloquies
because	bush	one	suitors
the	presidents	because	counterclaims
would	assailing	that	reprove
not	impeaching	they	emboldens
what	upbraided	same	rationales
when	alluded	well	overbid
also	barack	the	frontmen

Table 3: Top 10 nearest neighbors using simple averaging of embeddings and ALC.

even from single instances. Next we demonstrate that each of these individual instances can be treated as an observation in a hypothesis-testing framework. Before doing so, while readers may be satisfied about the performance of ALC in small samples, they may wonder about its performance in *large* samples. That is, whether it converges to the inferences one would make from a 'full' corpus model as the number of instances increases; the answer is 'yes' and we provide more details in Supporting Information C.

4.2 à la Carte on Text embedding regression model: `conText`

Recall the original statement of the relationship between the embedding of a focal word and the embeddings of the words within its context: $\mathbf{v}_w = \mathbf{A}\mathbb{E}[\mathbf{u}_w]$. Here we note that because the matrix $\mathbf{A}$ is constant we can easily swap it into the expectation and then calculate the resulting expectation conditional on some covariate X: $\mathbb{E}[\mathbf{A}\mathbf{u}_w|X]$. In particular, this can be done implicitly through a multivariate regression procedure. In the case of word meanings in discrete subgroups, this is exactly the same as the use of ALC applied above.

To illustrate our set up, suppose that each $\mathbf{v}_{w_i}$ is the embedding of a particular instance of a given word in some particular context, like `Trump`. Each is of some dimension, D and thus each 'observation' in this setting is a $1 \times D$ embedding vector. We can stack these to produce an outcome variable $\mathbf{Y}$ which is of dimensions n (the number of instances of a given word) by D. The usual multivariate matrix equation is then:

$$\underbrace{\mathbf{Y}}_{n\times D} = \underbrace{\mathbf{X}}_{n\times p+1} \underbrace{\boldsymbol{\beta}}_{p+1\times D} + \underbrace{\mathbf{E}}_{n\times D} \tag{2}$$

where $\mathbf{X}$ is a matrix of p covariates and includes a constant term, while $\boldsymbol{\beta}$ is a set of p coefficients and an intercept (all of dimension D). Then $\mathbf{E}$ is an error term.

To keep matters simple, suppose that there is a constant and then one binary covariate indicating group membership (in the group, or not). Then, the coefficient $\boldsymbol{\beta_0}$ (the first row of the matrix $\boldsymbol{\beta}$) is equivalent to averaging over all instances of the target word belonging to those not in the group. Meanwhile, $\boldsymbol{\beta_0} + \boldsymbol{\beta_1}$ (the second row of $\boldsymbol{\beta}$) is equivalent to averaging over all instances of the target word that belong to the group (i.e. for which the covariate takes the value 1, as opposed to zero). In the more general case of continuous covariates, this provides a model-based estimate of the embedding among all instances at a given level of the covariate space.

The key outputs from this à la Carte on Text (`conText`) embedding 'regression' model are:

- the coefficients themselves, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$. These can be used to calculate the estimated embeddings for the word in question. We can take the cosine distance between these implied embeddings and the (pre-trained) embeddings of other words to obtain the nearest neighbors for the two groups.

- the (Euclidean) norms of the coefficients. These will now be scalars (distances) rather than the vectors of the original coefficients. In the categorical covariate case, these tell us how different one group is to another in a *relative* sense. While the magnitude of this difference is not directly interpretable, we can nonetheless comment on whether it is statistically significantly different from zero. To do this, we use a variant of covariate assignment shuffling suggested by Gentzkow, Shapiro and Taddy (2019). In particular, we randomly shuffle the entries of the $\mathbf{Y}$ column and run the regression many (here

100) times. Each time, we record the norms of the coefficients. We then compute the proportion of those values that are larger than the *observed* norms (i.e. with the true group assignments). This is the empirical p-value.

Note that, if desired, one can obtain the sampling distribution (and thus standard errors) of the (normed) coefficients via non-parametric bootstrap. This allows for comments on the *relative* size of differences in embeddings across and within groups as defined by their covariates. We now show how the `conText` model may be used in a real estimation problem.

4.3 Our Framework in Action: Pre-Post Election Hypothesis Testing

We can compare the change in the usage of the word `Trump` to the change in the usage of the word `Clinton` after the 2016 election. Given Trump won the election and subsequently became President—a major break with respect to his real-estate/celebrity past—we expect a statistically significant change for `Trump` relative to any changes in the usage of `Clinton`.

We proceed as follows: for each target word-period combination—`Clinton` and `Trump`, pre-election (2011–2014) and post-election (2017-2020)— we embed each individual instance of the focal word from our *New York Times* corpus of leading article paragraphs, and estimate the following regression:

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{Trump} + \boldsymbol{\beta}_2 \text{Post\_Election} + \boldsymbol{\beta}_3 \text{Trump} \times \text{Post\_Election} + \mathbf{E} \tag{3}$$

where Trump is an indicator variable equal 1 for `Trump` instances, 0 otherwise. Likewise Post_Election is a dummy variable equal 1 for 2017-2020 instances of `Trump` or `Clinton`. As before, this is simply a regression-based estimator for the individual sub-groups. We will use permutation for hypothesis testing.

Figure 4 plots the norm of the $\hat{\boldsymbol{\beta}}$s. To reiterate, norming means the coefficient vectors become scalars. The significant positive value on the `Trump x Post_Election` coefficient indicates the expected additional shift in the usage of `Trump` post-election over and above the shift in the usage of `Clinton`.

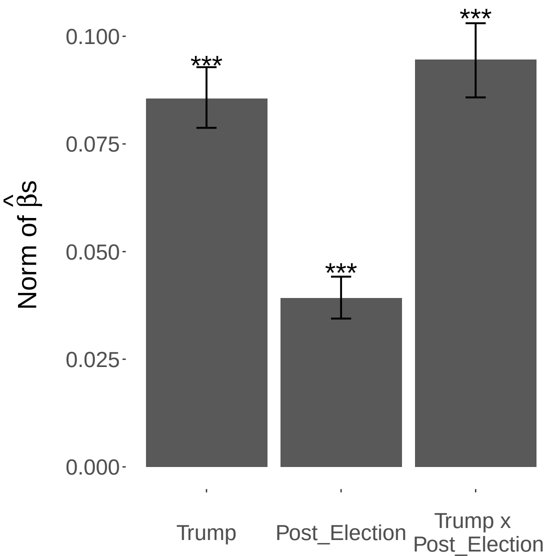


Figure 4: Relative semantic shift from `Trump`. Values are the norm of $\hat{\boldsymbol{\beta}}$ and bootstrap confidence intervals.

While this news is encouraging, readers may wonder how the `conText` regression model performs relative to a 'natural' alternative—specifically, a full embeddings model fit to each use of the term by covariate value(s). This would require the entire corpus (rather than just the instances of `Trump` and `Clinton`) and would be computationally slow, but perhaps it would yield more accurate inferences. As we demonstrate in Supporting Information D, inferences are similar and our approach is more stable by virtue of holding constant embeddings for all words but the focal word.

5 Results

We now turn to substantive use cases, beginning with partisan differences in the United States.

5.1 Partisanship, Ideology and Gender Differences

We want to evaluate partisan and gender differences in the usage of a given term in Congress–Sessions 111-114 (Obama years). Our focus is a set of target words known to be politically charged: `abortion`, `immigration` and `marriage`. We also include three non-partisan stopwords—`and`, `the` and `but`—in our target set as comparison.

We estimate the following multivariate multiple regression model for each of our words:

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\texttt{Republican} + \boldsymbol{\beta}_2\texttt{Male} + \mathbf{E}. \tag{4}$$

The dependent variable is an ALC embedding of each individual realization in the corpus. For the righthand side, we use indicator variables (Republican or otherwise; Male or otherwise). We use permutation to approximate the null and bootstrapping to quantify the sampling variance.

Note again that magnitudes have no natural absolute interpretation, but can be compared relatively: that is, a larger coefficient on X_i relative to X_j implies the difference in embeddings for the groups defined by i is larger than the difference in the groups as defined by j. Our actual results are displayed in Figure 5. The 'Male' coefficient is the average difference across the gender classes, controlling for party. The 'Republican' coefficient is the average difference across the parties, controlling for gender.

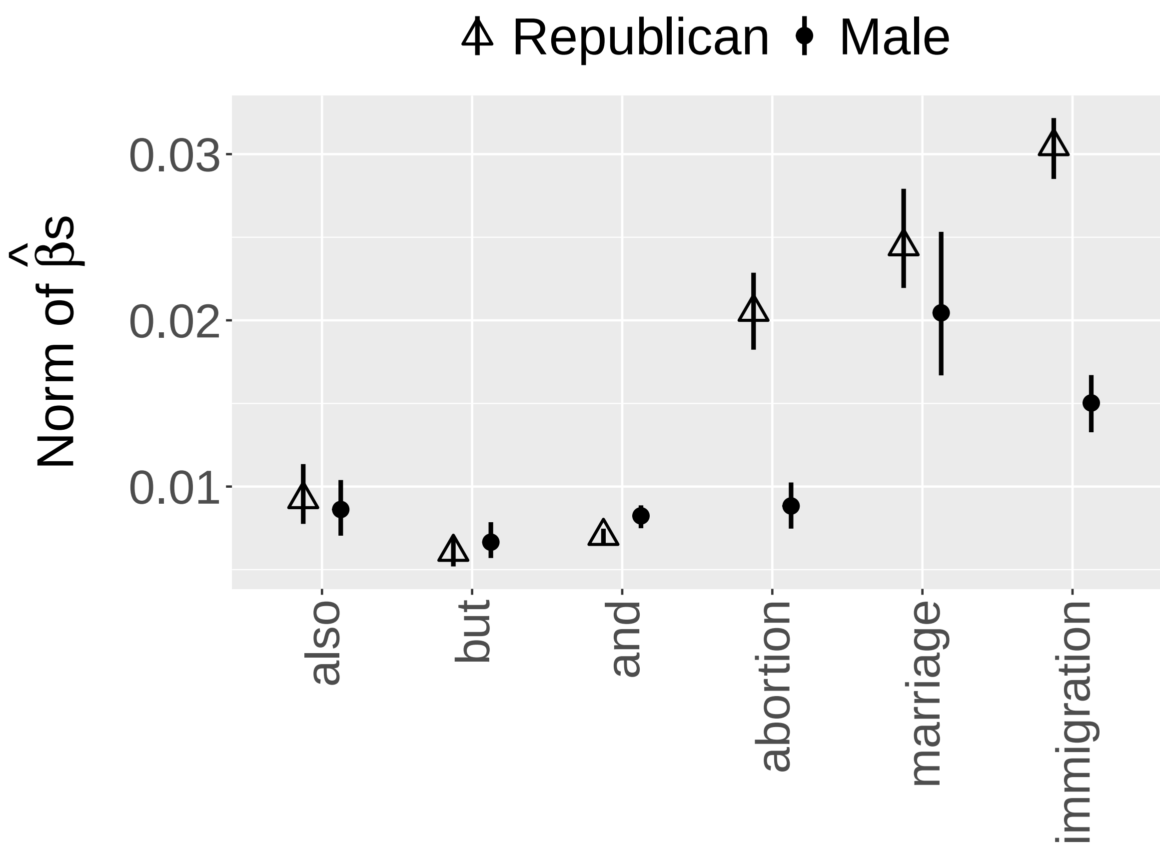


Figure 5: Differences in word meaning by gender and party: generally, different genders in the same party have more similar understanding of a term, than the same gender across parties.

As expected, the differences across parties and across genders, is much larger for the more political terms—relative to function words. But, in addition, embeddings differ more by party than they do by gender. That is, on average, men and women *within* a party have more similar understandings of the terms than men and women *across* parties.

The "most partisan" target in our set is `immigration`. Table 4 shows the top 10 nearest neighbors for each party. One reading of these nearest neighbors is that Democrats were pushing for reform of existing laws while Republicans were mainly arguing for enforcement. We can corroborate this via the top *nearest contexts*, i.e. the individual contexts of `immigration`—embedded using ALC—that are closest to each party's ALC embedding of the term (see Table 5). This suggests some validity of our general approach.

Democrats	enact, overhauling, reform, legislation, enacting, overhaul, reforming, revamp, entitlement, bipartisan
Republicans	enforce, laws, enact, enacting, legislate, legislations, enforcing, regularize, immigration, legislation

Table 4: Top 10 nearest neighbors for the target term `immigration`.

Democrats
this congress to take on comprehensive `immigration` reform and fix our broken immigration
should get to work on comprehensive `immigration` reform the immigration system we have
Republicans
administration wants to ignore our nations `immigration` laws and immigration process the problem
we need true reform of our `immigration` laws starting with border security and

Table 5: Subset of top nearest contexts for the target term `immigration`.

Our approach is not limited to binary covariates. To illustrate, we regress the target word `immigration` on the first dimension of the NOMINATE score[9]—understood to capture the Liberal-Conservative spectrum on economic matters (Poole, 2005). This approximates a whole sequence of separate embeddings for each speaker, approximated using a line in the NOMINATE space. We estimate the following regression:

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{NOMINATE} + \mathbf{E} \tag{5}$$

We next predict an ALC embedding for `immigration` at each percentile of the NOMINATE score and compute its cosine similarity with a small set of hand-picked features. Figure 6 plots these results. Consistent with our results above, we observe how the predicted ALC

[9]Downloaded from `https://voteview.com/data`.

embedding for `immigration` is closer to `enforce` and `illegals` at higher values of the NOMINATE score. It is closer to `reform` and `bipartisan` at lower values. The feature `amend` on the other hand, shows similar values across the full range.

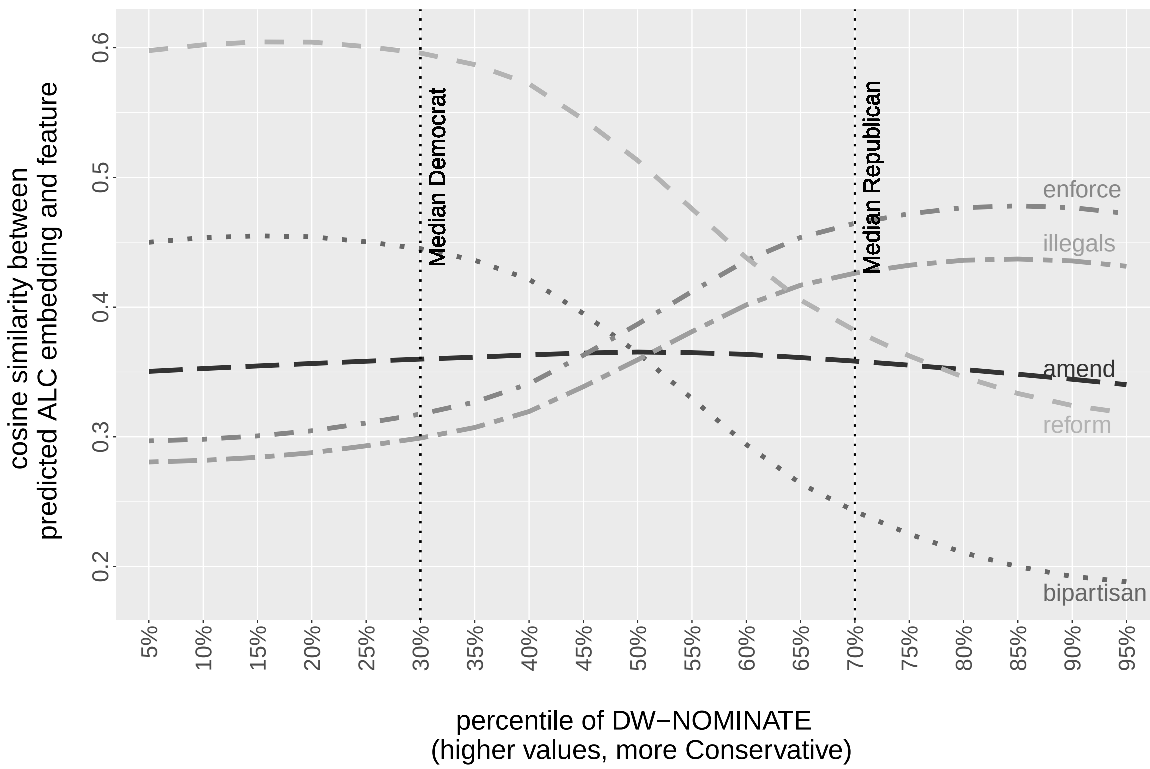


Figure 6: Cosine similarity (loess smoothed) between various words and "immigration" at each percentile of NOMINATE scores. We mark the median Democrat and median Republican to help calibrate the scale.

5.2 The Meaning of 'Empire'

Recall that our plan for the second case-study was to compare the embedding of `Empire` in the UK and US context for the period 1935–2010. In the estimation we use the top (most frequent) 5000 tokens of the combined corpora and we estimate a 300-dimensional `GloVe` model and corresponding **A** matrix specific to the corpus. The multivariate regression analogy is

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \texttt{Congressional Record} + \mathbf{E} \quad (6)$$

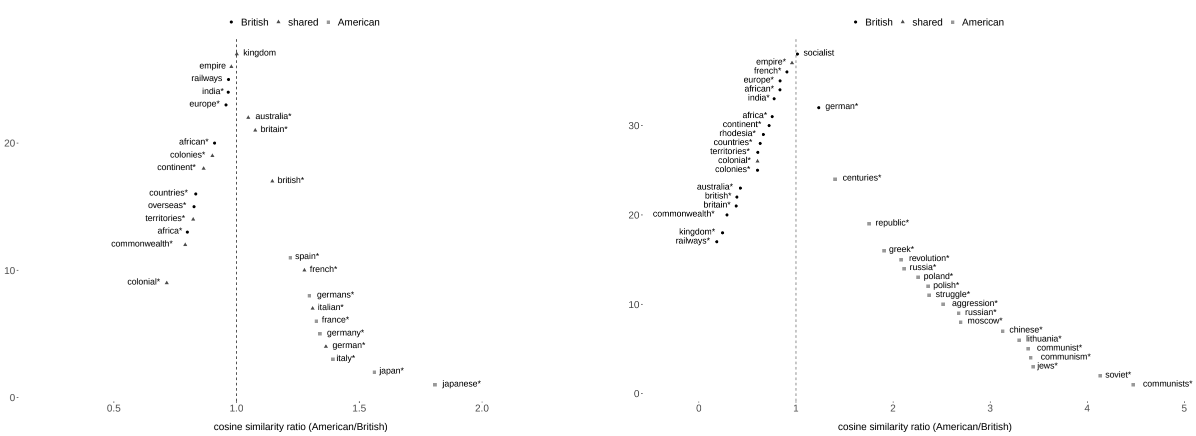
estimated for every year of the period. Interest focuses on the (normed) value of β_1: when this rises, the use of `Empire` is becoming less similar across the corpora (Congress is becoming more distinctive). The time series of the β_1s is given in Figure 7. The basic summary is that, sometime around 1947-48, there was a once-and-for-all increase in the distance between US and UK understandings of `Empire`. We confirmed this with a structural break test (Bai and Perron, 1998).

Figure 7: Norm of the British and American difference in understanding of `Empire`, 1935–2010: larger values imply the uses are more different.

To understand the substance of the change, consider Figure 8. We report the 'most American' and 'most British' (with reference to the parliaments) terms from the period either side of the split in the series. Specifically, we calculate the cosine similarity between the ALC embedding for `Empire` and each nearest neighbor in the UK and US corpus. The x-axis is the *ratio* of these similarities: when it is large, the word is relatively closer to the US understanding of `Empire` than to the UK one. An asterisk by the term implies that ratio's deviation from 1 is statistically significantly larger than its permuted value, $p < 0.01$.

The y-axis reports the rank of the word in terms of distance from 0: words near the bottom of the plot are more distinct than those near the top.

The main observation is that in the pre-period, British and American legislators talk about `Empire` primarily in connection with the old European powers: e.g. Britain and France. By contrast, the vocabularies are radically different in the post-break period. While the UK parliament continues to talk of the "British" empire (and its travails in "India" and "Rhodesia"), the US focus has switched. For the Americans, understandings of empire are specifically with respect Soviet imperial ambitions and we see this in the most distinct nearest neighbors "invasion", "Soviet" and "communists", with explicit references to eastern European nations like "Lithuania".



(a) Pre-1948

(b) Post-1948

Figure 8: UK and US discussions of `Empire` diverged after 1948: most US and UK nearest neighbors pre and post estimated breakpoint.

5.3 Brexit Sentiment from the Backbenches

Our goal is to estimate the sentiment of the Conservative party towards the EU in the House of Commons. First, the underlying debate text and meta-data is from Osnabrügge, Hobolt and Rodon (2021), covering the period 2001–2019. We are interested in both major parties of government, Labour and Conservatives. We divide those parties' MPs by role: Cabinet (or Shadow Cabinet in opposition) members of the government party are 'cabinet', all others are 'backbenchers', by definition. We compare policy sentiment in three areas: education (where our term of interest is 'education'), health ('nhs') and the EU ('eu').

In what follows, each observation for us is a representation of the sentiment of a party-rank-month triple towards a given term. For instance, (the average) Conservative-backbencher-July 2015 sentiment towards "health". We describe our approach in SI E; in essence we measure the inner product between the term of interest to the aggregate embeddings of the (positive and negative) words from a sentiment dictionary (Warriner, Kuperman and Brysbaert, 2013). We then rescale within party, obtaining Figure 9. There, each column is a policy area: education, health and then the EU. The rows represent the Conservatives at the top, and Labour at the bottom, with the correlation between Tory backbenchers and cabinet in the middle. We see an obvious "government versus opposition" Westminster dynamic: when Labour is in power (so, from the start of the data to 2010), Labour leaders and backbenchers are generally enthusiastic about government policy. That is, their valence is mostly positive, which makes sense given almost total government agenda control (i.e. the policy being discussed is government policy). The Conservatives are the converse: both elites and backbenchers have negative valence for government policy when in opposition, but are much more enthusiastic when in government. This is true for education, and health to a lesser extent. So far, so expected.

But the subject of the EU (the "eu" column) is different (top right cell). We see that even after the Conservatives come to power (marked by the broken black line in 2010) backbench opinion on government policy towards Europe is negative. By contrast, the Tory leadership are positive about their own policy on this subject. Only after the Conservatives introduce referendum legislation (the broken vertical line in 2015) upon winning the General Election, do the backbenchers begin to trend positive towards government policy. The middle row makes this more explicit: the correlation between Tory leadership and backbench sentiment is approximately zero for education and health, but *negative* for the EU—i.e. moving in opposite directions. Our finding here is that Cameron never convinced the average

Conservative backbencher that his EU policy was something about which they should feel positive.

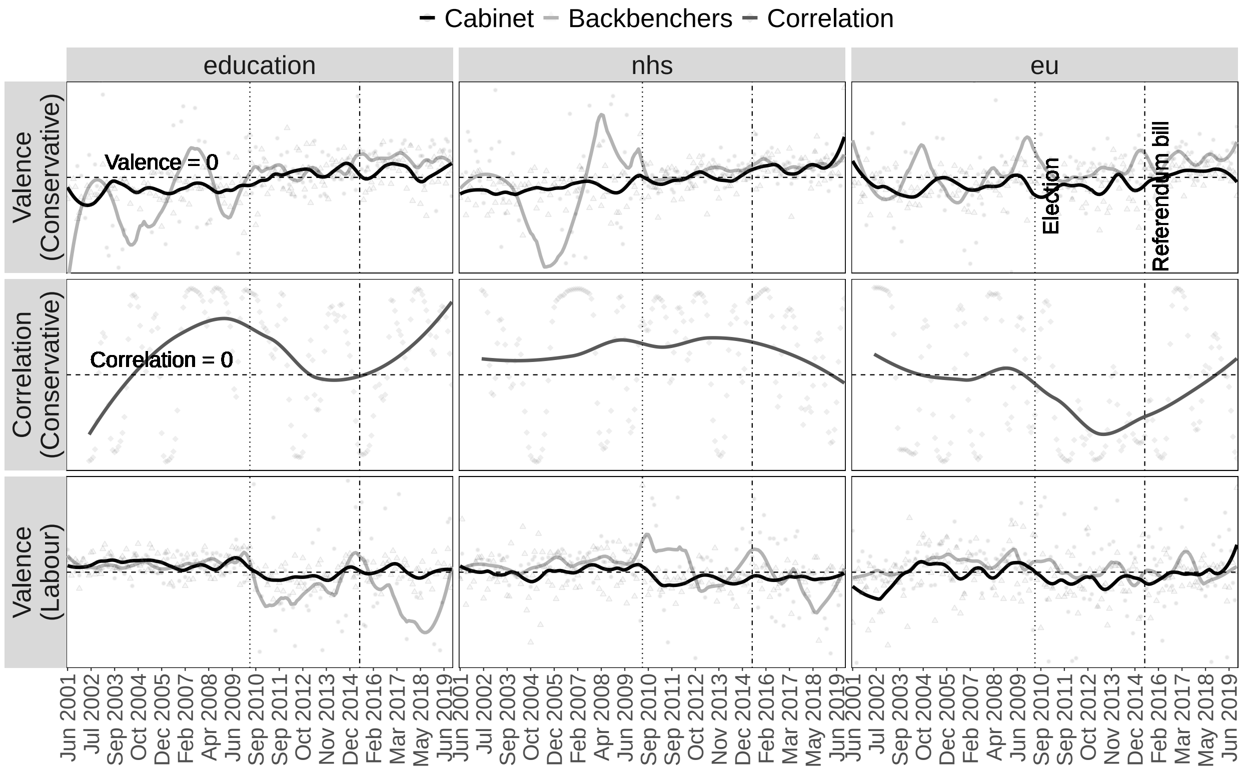


Figure 9: Conservative backbenchers were unsatisfied with their own government's EU policy prior to the referendum. Each column of the plot is a policy area (with the seed word used to calculate sentiment). Those areas are: education (`education`), health (`nhs`) and the EU (`eu`). Note the middle-right plot: rank-and-file Conservative MP sentiment on EU policy are negatively correlated with their leadership's sentiment.

A more traditional approach would be to count the number of occurrences of terms in the sentiment dictionary and assign each speech a net valence score. Figure 10 displays that result. Patterns are harder to read. More importantly, only 56% of the terms in the dictionary occur in the speeches and a full 68% of speeches had no overlap with the set of dictionary terms—and thus receive a score of 0. This contrasts with the 99% of terms in the dictionary appearing in the pre-trained embeddings, allowing for all speeches to be scored. This is due to the continuity of the embedding space.

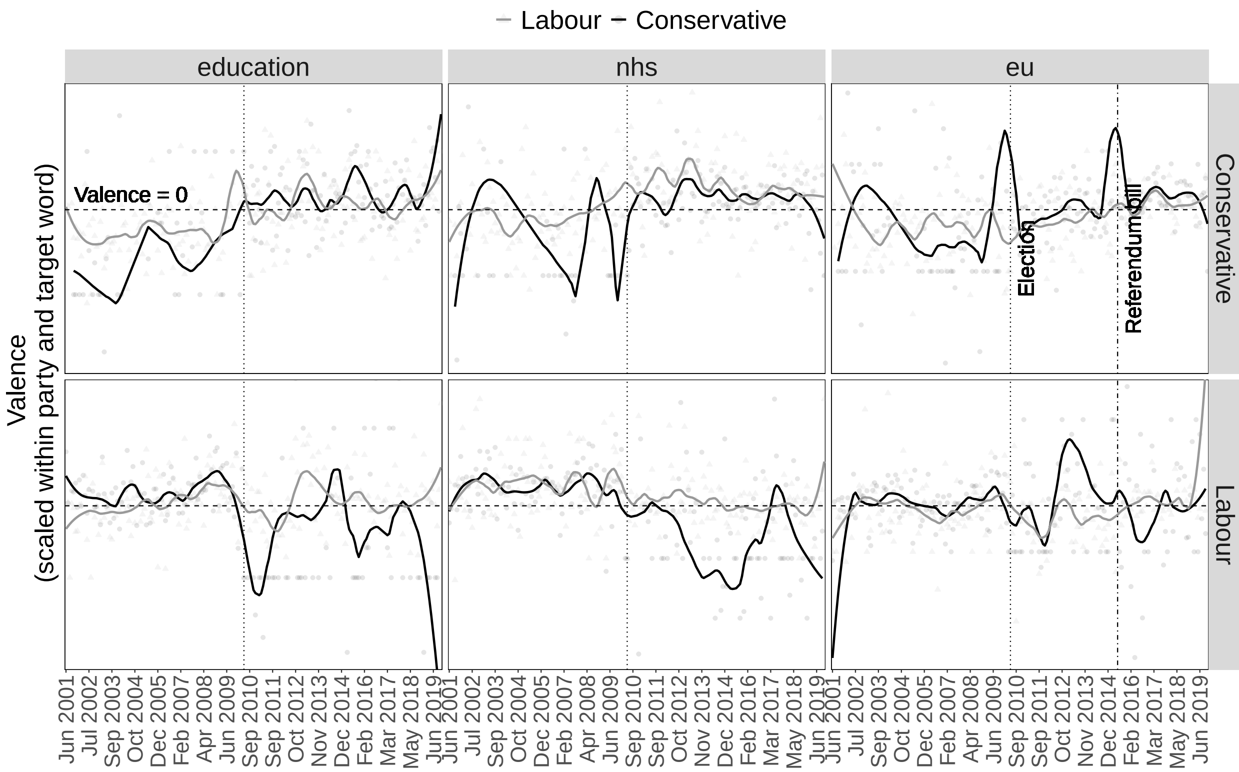


Figure 10: Replication of Figure 9 using a dictionary approach. The sentiment patterns are less obvious.

6 Advice to Practitioners: Experiments, Limitations, Challenges

Our approach requires no active tuning of parameters, but that does not mean that there are no *choices* to make. For example, the end-user can opt for different context window sizes (literally, the number of words either side of the target word), as well as different preprocessing regimes. To guide practice, we now summarize experiments we did on real texts. Below, we use 'pre-trained' to refer to embeddings that have been fit to some large (typically on-line) data collection like Wikipedia. We use 'locally fit' to mean embeddings produced from—i.e. vectors learned from—the texts one is studying (e.g. Congressional debates). We note that Rodriguez and Spirling (2022) provide extensive results on this comparison for current models; thus here we are mostly extending those enquiries to our specific approach. Our full write up can be seen in Supporting Information F–H. The following are the most important results.

First, we conducted a series of *supervised* tasks, where the aim is to separate the uses of `trump` vs `Trump` per our example in Section 4.1. We found that removing stopwords and using bigger context windows results in marginally better performance. That is, if the researcher's goal is to differentiate two separate uses of a term (or something related, such as classifying documents), more data—i.e. larger contexts, less noise—makes sense. To be candid though, we do not think such a task—where the goal is a version of accuracy—is a particularly common one in political science.

We contend a more common is seeking high quality embeddings *per se*. That is, vector representations of terms that correctly capture the 'true' embedding (low bias) and are simultaneously consistent across similar specifications (low variance, in terms of model choices). We give many more details in the SI, but the basic idea here is to fit locally trained embeddings—with context window size 2, 6 and 12—to the *Congressional Record* corpus (Sessions 107–114). We then treat those embeddings as targets to be recovered from various ALC-based models that follow, with closer approximations being deemed better. As an additional 'ground truth' model we use Stanford `GloVe` pre-trained embeddings (window size 6, 300 dimensions). We narrow our comparisons to a set of 'political' terms as given by Rodriguez and Spirling (2022). We have five lessons from our experiments:

1. **Pretraining and windows**: given a large corpus, local training of a full embeddings model and corresponding $\mathbf{A}$ matrix makes sense. Our suggested approach can then be used to cheaply and flexibly study differences across groups. Barring that, using pre-trained embeddings trained on large online corpora (e.g. Stanford `GloVe`) provides a very reasonable approximation which can be further improved by estimating an $\mathbf{A}$ matrix specific to the local corpus. But again, if data is scarce, using an $\mathbf{A}$ matrix trained on the original online corpus (e.g. Khodak et al. (2018)'s $\mathbf{A}$ in the case of GloVe) leads to very reasonable results. In terms of context window size, avoid small

windows (of size < 5). Windows of size 6 and 12 perform very similarly to each other, and acceptably well in an absolute sense.

2. **Preprocessing**: removing stopwords from contexts used in estimating ALC embeddings makes very little difference to any type of performance. In general, apply the same preprocessing to the ALC contexts as was applied at the stage of estimating the embeddings and **A** matrix—e.g. if stopwords were not removed, then do not remove stopwords. Stemming/lemmatization does not change results much in practice.

3. **Similarity metrics:** the conventional cosine similarity provides interpretable neighbors, but the inner product often delivers very similar results.

4. **Uncertainty:** uncertainty in the calculation of the **A** matrix is minimal and unlikely to be consequential for topline results.

5. **Changing contexts over time:** potential changes to contexts of targets is a second-order concern, at least for texts from the past 100 years or so.

Before concluding, we note that as with almost all descriptive techniques, the ultimate substantive interpretation of the findings is left with the researcher to validate. It is hard to give general advice on how this might be done, so we refer readers to two approaches. First, one can try to triangulate using various types of validity: semantic, convergent construct, predictive and so on (see, Quinn et al., 2010, for discussion). Second, crowd-sourced validation methods may be appropriate (see Rodriguez and Spirling, 2022; Ying, Montgomery and Stewart, 2021).

Finally, we alert readers to the fact that all of our analyses can be implemented using the `conText` software package in `R` (see Supporting Information I and `https://github.com/prodriguezsosa/conText`).

7 Conclusion

"Contextomy"—the art of quoting out of context to ensure that a speaker is misrepresented—has a long and troubling history in politics (McGlone, 2005). It works because judicious removal of surrounding text can so quickly alter how audiences perceive a central message. Understanding how context affects meaning is thus of profound interest in our polarized times. But it is difficult—to measure and model. This is especially true in politics, where our corpora may be small and our term counts low. Yet we simultaneously want statistical machinery that allows us to speak of statistically significant effects of covariates. This paper begins to address these problems.

Specifically, we proposed a flexible approach to study differences in semantics between groups and over time using high-quality pre-trained embeddings: the `conText` embedding regression model. It has advantages over previous efforts, and that it can reveal new things about politics. We explained how controversial terms divide parties not simply in the way they are attached to topics of debate, but in their very meaning. Similarly, we showed that understandings of terms like "empire" are not fixed, even in the relatively short-term, and instead develop in-line with interests in international relations. We showed that our approach can be used to measure sentiment towards policy. It is not hard to imagine other applications. For example, there is evidence that voters prefer broad-based appeals (Hersh and Schaffner, 2013), but these are only possible in cases where meanings are sufficiently similar within groups. Our technique could be used to explore this tension. Similarly, what is deemed the "correct" interpretation of treaties (e.g. Simmons, 2010) or constitutions matters. Our methods could help structure studies of these changes.

We built our framework on the ALC embedding strategy. But our general approach is not inextricably connected to this particular method for estimating contextually specific meanings. We used it because it is transparent, efficient, and computationally simple. We introduced a regression framework for understanding word meanings using individual instance embeddings as observations. This may be easily extended to more complex functional forms.

There are many potential directions for the framework; we highlight two. First, ALC assumes that the meaning of non-focal words is essentially constant. This first-order approximation could be extended with second-order information—which words co-occur with words that co-occur with the focal words—but it is unclear how much meaning would have to change across groups for this to matter. Second, we are estimating means in high-dimensions using only a few data points. This is always difficult (see Gentzkow, Shapiro and Taddy, 2019) and our estimates of the norms have a finite-sample bias for rare words. Thus care is needed when comparing words or groups with substantially different amounts of available data. Future work could consider the role of term frequency in these measures of meaning.

As social scientists develop further methods to study these problems, this will sharpen questions which will in turn spur better methods. But to reiterate, technical machinery cannot, of itself, answer substantive questions. That is, claims about meaning must be validated, and the way that differences in measured quantities are interpreted will always be subject to debate. We hope that the `conText` model that we have laid out here can provide a useful foundation for future work.

References

Antoniak, Maria and David Mimno. 2018. "Evaluating the stability of embedding-based word similarities." *Transactions of the Association for Computational Linguistics* 6:107–119.

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma and Andrej Risteski. 2016. "A latent variable model approach to pmi-based word embeddings." *Transactions of the Association for Computational Linguistics* 4:385–399.

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma and Andrej Risteski. 2018. "Linear algebraic structure of word senses, with applications to polysemy." *Transactions of the Association for Computational Linguistics* 6:483–495.

Aslett, Kevin, Nora Webb Williams, Andreu Casas, Wesley Zuidema and John Wilkerson. 2022. "What Was the Problem in Parkland? Using Social Media to Measure the Effectiveness of Issue Frames." *Policy Studies Journal* 50(1):266–289.

Austin, John Langshaw. 1962. *How to do things with words.* Oxford: Oxford University Press.

Bai, Jushan and Pierre Perron. 1998. "Estimating and Testing Linear Models with Multiple Structural Changes." *Econometrica* 66(1):47–78.

Caliskan, Aylin, Joanna J Bryson and Arvind Narayanan. 2017. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356(6334):183–186.

Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy and Noah A. Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Denver, Colorado: Association for Computational Linguistics pp. 1606–1615.

Firth, John Rupert. 1957. *Studies in linguistic analysis.* Wiley-Blackwell.

Ford, Robert and Matthew Goodwin. 2017. "Britain after Brexit: A nation divided." *Journal* *of Democracy* 28(1):17–30.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky and James Zou. 2018. "Word embeddings quantify 100 years of gender and ethnic stereotypes." *Proceedings of the National Academy of Sciences* 115(16):E3635–E3644.

Geertz, Clifford. 1973. "Thick description: Toward an interpretive theory of culture." *Turning points in qualitative research: Tying knots in a handkerchief* 3:143–168.

Gentzkow, Matthew, Jesse M Shapiro and Matt Taddy. 2019. "Measuring group differences in high-dimensional choices: method and application to congressional speech." *Econometrica* 87(4):1307–1340.

Gentzkow, Matthew, J.M. Shapiro and Matt Taddy. 2018. "Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts.".

Grimmer, Justin. 2010. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." *Political Analysis* 18(1):1–35.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21(3):267–297.

Hamilton, William L., Jure Leskovec and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics pp. 1489–1501.

Han, Rujun, Michael Gill, Arthur Spirling and Kyunghyun Cho. 2018. Conditional word embedding and hypothesis testing via bayes-by-backprop. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. pp. 4890–4895.

Harris, Zellig S. 1954. "Distributional structure." *Word* 10(2-3):146–162.

Hennessy, Peter. 1992. *Never Again: Britain 1945-1951*. Penguin UK.

Hersh, Eitan D and Brian F Schaffner. 2013. "Targeted campaign appeals and the value of ambiguity." *The Journal of Politics* 75(2):520–534.

Hinton, Geoffrey E. 1986. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*. Vol. 1 Amherst, MA p. 12.

Hobolt, Sara B. 2016. "The Brexit vote: a divided nation, a divided continent." *Journal of European Public Policy* 23(9):1259–1277.

Hopkins, Daniel J. 2018. "The exaggerated life of death panels? The limited but real influence of elite rhetoric in the 2009–2010 health care debate." *Political Behavior* 40(3):681–709.

Khodak, Mikhail, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon M. Stewart and Sanjeev Arora. 2018. A la carte embedding: Cheap but effective induction of semantic feature vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne, Australia: Association for Computational Linguistics pp. 12–22.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, MD, USA: Association for Computational Linguistics pp. 61–65.

Kozlowski, Austin C, Matt Taddy and James A Evans. 2019. "The geometry of culture: Analyzing the meanings of class through word embeddings." *American Sociological Review* 84(5):905–949.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*. pp. 625–635.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski and Erik Velldal. 2018. "Diachronic word embeddings and semantic shifts: a survey." *arXiv preprint arXiv:1806.03537* .

Lauretig, Adam. 2019. Identification, Interpretability, and Bayesian Word Embeddings. In *Proceedings of the Third Workshop on Natural Language Processing and Computational*

Social Science. Minneapolis, Minnesota: Association for Computational Linguistics pp. 7–17.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. 2019. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* .

McGlone, Matthew S. 2005. "Contextomy: the art of quoting out of context." *Media, Culture & Society* 27(4):511–522.

Mikolov, Tomás, Ilya Sutskever, Kai Chen, Greg S. Corrado and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, ed. C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger. Curran Associates, Inc. pp. 3111–3119.

Miller, George A and Walter G Charles. 1991. "Contextual correlates of semantic similarity." *Language and cognitive processes* 6(1):1–28.

Monroe, Burt L, Michael P Colaresi and Kevin M Quinn. 2008. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16(4):372–403.

Osnabrügge, Moritz, Sara B Hobolt and Toni Rodon. 2021. "Playing to the gallery: Emotive rhetoric in parliaments." *American Political Science Review* 115(3):885–899.

Park, Baekkwan, Kevin Greene and Michael Colaresi. 2020. "Human rights are (increasingly) plural: Learning the changing taxonomy of human rights from large-scale text reveals information effects." *American Political Science Review* 114(3):888–910.

Pennington, Jeffrey, Richard Socher and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. pp. 1532–1543.

Poole, Keith T. 2005. *Spatial models of parliamentary voting*. Cambridge University Press.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Rheault, Ludovic and Christopher Cochrane. 2020. "Word embeddings for the analysis of ideological placement in parliamentary corpora." *Political Analysis* 28(1):112–133.

Roberts, Margaret E, Brandon M Stewart and Edoardo M Airoldi. 2016. "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association* 111(515):988–1003.

Rodman, Emma. 2020. "A timely intervention: Tracking the changing meanings of political concepts with word vectors." *Political Analysis* 28(1):87–111.

Rodriguez, Pedro L. and Arthur Spirling. 2022. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *The Journal of Politics* 84(1):101–115.

Rudolph, Maja, Francisco Ruiz, Susan Athey and David Blei. 2017. Structured embedding models for grouped data. In *Advances in Neural Information Processing Systems*. pp. 251–261.

Simmons, Beth. 2010. "Treaty compliance and violation." *Annual Review of Political Science* 13:273–296.

Skinner, Quentin. 1969. "Meaning and Understanding in the History of Ideas." *History and theory* 8(1):3–53.

Slapin, Jonathan B, Justin H Kirkland, Joseph A Lazzaro, Patrick A Leslie and Tom O'grady. 2018. "Ideology, grandstanding, and strategic party disloyalty in the British Parliament." *American Political Science Review* 112(1):15–30.

Tversky, Amos and Daniel Kahneman. 1981. "The framing of decisions and the psychology of choice." *Science* 211(4481):453–458.

Verba, Sidney and Gabriel Almond. 1963. *The civic culture: Political attitudes and democracy in five nations*. Princeton, NJ: Princeton University Press.

Warriner, Amy Beth, Victor Kuperman and Marc Brysbaert. 2013. "Norms of valence, arousal, and dominance for 13,915 English lemmas." *Behavior research methods* 45(4):1191–1207.

Wu, Patrick Y, Walter R Mebane Jr, Logan Woods, Joseph Klaver and Preston Due. 2019. "Partisan Associations of Twitter Users Based on Their Self-descriptions and Word Embeddings.".

Yin, Zi, Vin Sachidananda and Balaji Prabhakar. 2018. The Global Anchor Method for Quantifying Linguistic Shifts and Domain Adaptation. In *Advances in Neural Information Processing Systems*, ed. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett. Vol. 31 Curran Associates, Inc.

Ying, Luwei, Jacob M Montgomery and Brandon M Stewart. 2021. "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures." *Political Analysis* pp. 1–20.

Online Supporting Information:

Embedding Regression: Models for Context-Specific Description and Inference

Contents (Appendix)

A Rodman: Details on Sample Sizes

A key challenge in Rodman's (2019) approach is that there is relatively little data (per time slice) to estimate embeddings from. Table A.1 presents the number of instances of each theme word for each period. Note that in almost 30% of the word-era combinations, there are fewer than 10 observations. Producing meaningful embeddings given these sample sizes is generally difficult.

	1855–	1880–	1905–	1930–	1955–	1980–	2005–
african_american	63	27	79	171	274	45	22
gender	4	41	374	560	460	258	284
german	1	2	62	512	13	2	2
race	5	15	76	188	190	34	38
treaty	3	1	143	216	30	3	1
Total Documents	80	102	496	1137	660	259	371

Table A.1: Number of instances of each category word in the Rodman corpus by 25 year time slice. All documents have the word `equality`. Many of the counts are quite low leading to a serious challenge for word embeddings.

B The Presidential Transition in Meaning

The meaning of `Trump`, the surname, underwent a transformation once Donald J. Trump was elected president of the United States in November 2016. This is a difficult case since the person being referred to is still the same entity, even though the meaning has shifted.

Using ALC, we embed a random sample of 100 mentions of `Trump` from 2001–2014 and 2017–2020, which we label celebrity `Trump` and president `Trump`, respectively. We do the same two cluster routine as above and inspect the 10 nearest neighbors—these are given in Table B.2. As we would expect, `Trump` in 2001–2014 is mentioned in the context of casinos and real-estate terms while `Trump` in 2017–2020 is mentioned in the context of terms associated with his presidency.

celebrity Trump	trump, ivanka, ivana, wynn, donald, casino, casinos, resorts, taj, caesars
President Trump	president, assailing, clinton, bush, impeach, impeachment, presidential, impeached, appointee

Table B.2: Top 10 nearest neighbors of the transformed cluster centroids. Top row (unshaded) is 2001–2014. Bottom row (shaded) is 2017–2020.

In Figure B.1 label the mentions of celebrity `Trump` and president `Trump`, respectively (results projected down to two-dimensions for visualization purposes). While the two groups overlap, as would be expected given mentions are all of the same person, it is clear mentions of `Trump` tend to cluster by period.

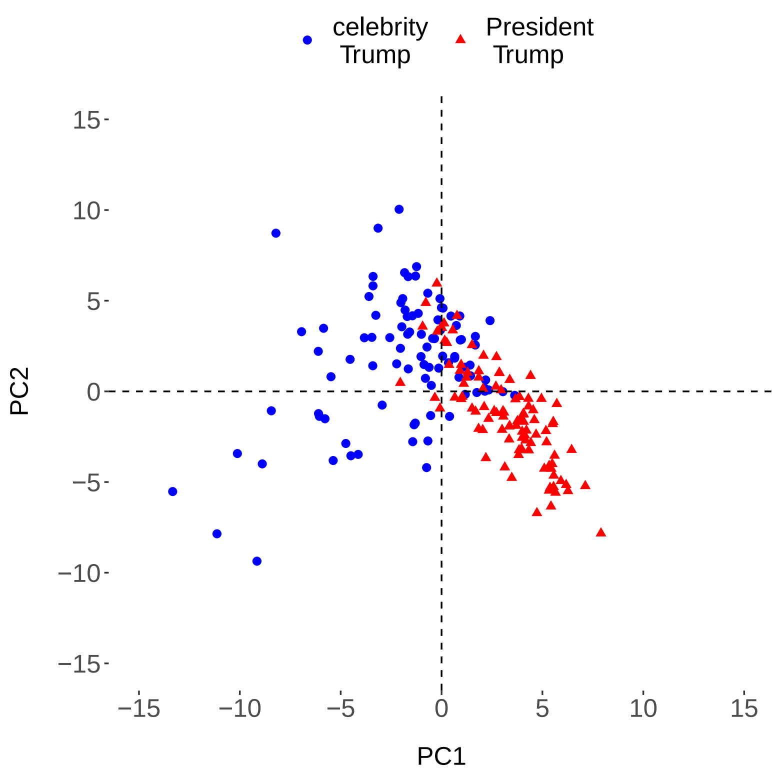


Figure B.1: Each observation represents a single realization of a context. Contexts for celebrity `Trump` include mentions of `Trump` in the New York Times during the period 2001-2014, while contexts for President `Trump` include mentions of `Trump` in the New York Times during the period 2017-2020.

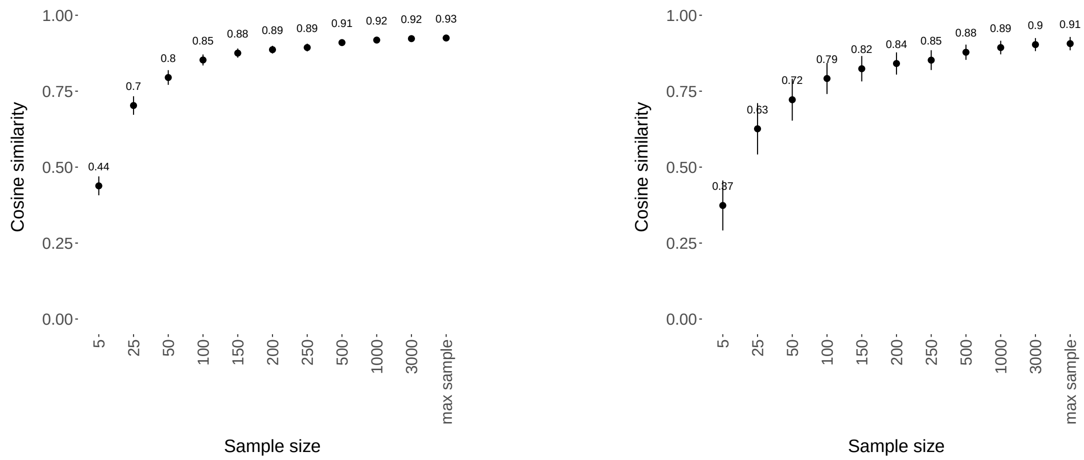
C Asymptotic Behavior

In this exercise we evaluate the asymptotic performance of our approach. That is, we want to know whether—and how quickly—ALC embeddings converge to embeddings from a fully trained, full corpus `GloVe` model, as we increase the number of instances ALC has access to. Obviously, we would hope that as the sample approaches the whole corpus, ALC 'looks like' a full corpus model.

For our corpus we use the *Congressional Record*. We begin by estimating a full `GloVe` embeddings model and a corresponding transformation matrix $\mathbf{A}$. Next we select a set of 20 target words from the corpus vocabulary, including 10 politics terms and 10 randomly sampled terms, and estimate their corresponding ALC embeddings. We vary the number of instances, from 5 to the total number of instances of each term.[10] Finally, we compute the cosine similarity between each ALC embedding and its corresponding embedding in the full `GloVe` model. Figure C.2 plots the results separately for the politics and random set of terms. We see that for both sets the ALC embeddings quickly converge to within a margin of error of the `GloVe` embeddings as the number of instances used to estimate the ALC embedding increases. This is expected and welcome behavior. In the case of the politics terms, with as few as fifty instances we see an average cosine similarity value of 0.8.[11]

[10] The set of politics terms are: `democracy`, `freedom`, `equality`, `justice`, `immigration`, `abortion`, `welfare`, `taxes`, `republican` and `democrat`. The set of random terms are: `adopt`, `appreciate`, `deserve`, `governments`, `however`, `insert`, `proposals`, `reduceds`, `temporary` and `thus`.

[11] Note, we do not expect this value to converge fully to 1 as the transformation matrix A is itself a regression estimate.



(a) Mean over 10 politics terms (b) Mean over 10 random terms

Figure C.2: Cosine similarity between a full `GloVe` (full corpus) embeddings model and ALC as a function of sample size.

D Benchmarking Embedding Regression against 'full' embeddings

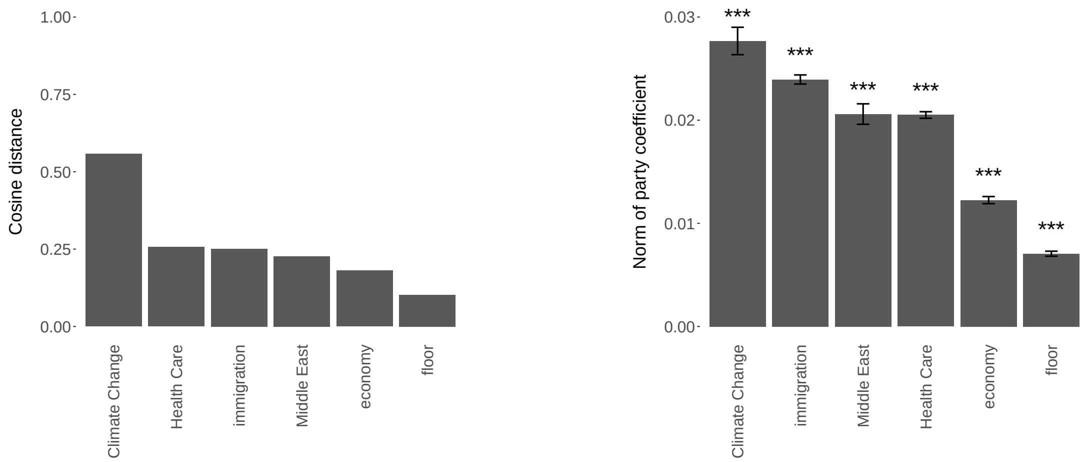
An alternative to our *regression* approach to quantifying group differences is to estimate a full `GloVe` embeddings model for each group's use of a term. For any given word this can be done by tagging (literally, slightly altering) the word in the corpus such that it appears differently for each different group. Estimating a full `GloVe` model on this tagged corpus yields group-specific embeddings for the tagged words. We can then use these embeddings to quantify group differences. This is computationally costly but provides us with a straightforward benchmark for our approach. Specifically we are interested in comparing inferences when applying both approaches to the following task: ranking a set of terms according to partisanship (in use).

For this exercise we use the Congressional Record corpus, sessions 111^{th}–114^{th} (the Obama years). As target words we use: `immigration`, `economy`, `climatechange`, `healthcare`, `middleeast` and, as a non-political control word we use `floor`.[12]

We tag every instance of a target word in the corpus with the party of its corresponding speaker, so for example, given a particular instance of `immigration` in a speech, we replace it with `immigrationd` if the author of the speech is a Democrat and with `immigrationr` if the author is a Republican. Given party specific embeddings for each target word we quantify partisanship using cosine distance, the higher the cosine distance, the more partisan the term. To quantify partisanship using our preferred approach we simply run a regression with party as a covariate and compute the norm of the resulting coefficient, the higher the norm of the party coefficient, the more partisan the term.

Figure D.3 plots both sets of results. Broadly speaking, the inferences one would draw from each are similar. On the one hand, `Climate Change` is clearly the most partisan issue while, as expected, our control term `floor` is the least partisan according to both models. `economy` stands out as the second least partisan according to both models. The remaining terms are similarly ranked except our approach suggests `immigration` is somewhat more partisan than `Health Care` and `Middle East`. All in all, the inferences from both approaches are not wildly different. In contrast to estimating a full `GloVe` embeddings model however, our approach is much faster, more stable—the solution does not vary across runs—and allows us to speak to the significance and sampling variance of our estimates.

[12]In the corpus we replace any mentions of `middle east` with `middleeast`, `health care` with `healthcare`, `immigrants` and `immigrant` with `immigration` and `climate change` with `climate change`.



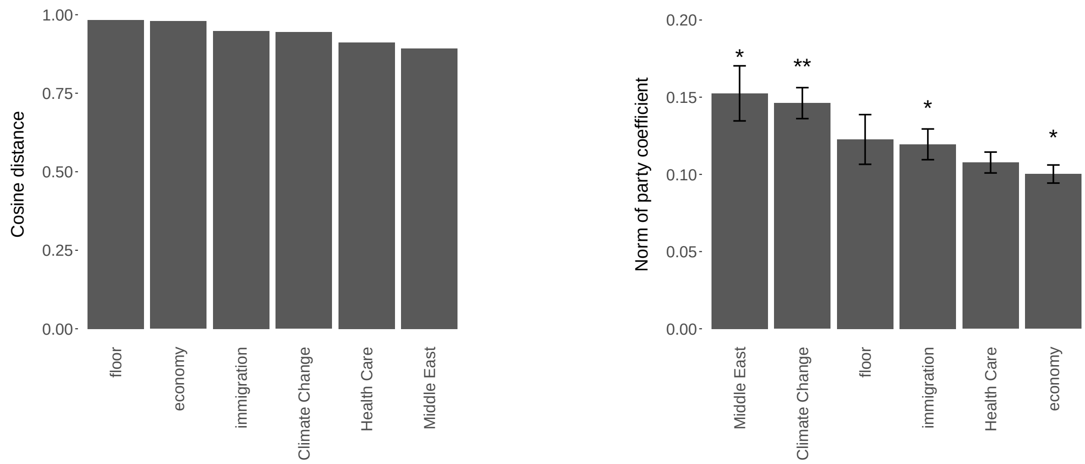
(a) `GloVe` ‘regression’

(b) `conText` regression

Figure D.3: Partisan differences using the Congressional Record corpus (Sessions 111$^{\text{th}}$ - 114$^{\text{th}}$).

Next we compare each model's performance with a significantly reduced sample, specifically one in which each target word appears in no more than five documents.[13] Our goal with this exercise is to compare how both methods fare in a small-sample world, relative to inferences using the full corpus. Figure D.4 plots both sets of results. In the case of the full GloVe model we see results are now flipped, with `floor` and `economy` showing the largest partisan differences. In contrast, the ALC results are comparable to the full-sample case. While `floor` shows a larger norm, it is not significant, and `Climate Change` remains the most significantly partisan of the target words. Combined, these results serve to highlight the added value of our approach, yielding similar inferences as the full embeddings model at a fraction of the cost and more robust in small-sample scenarios.

[13]To build this corpus we identify for each target word all documents containing the word and randomly sample five of these. We exclude from this sample any document containing multiple target words. Documents that do not contain any of the target words remain part of the corpus.



(a) GloVe (b) ALC

Figure D.4: Partisan differences using the Congressional Record corpus (Sessions 111$^{\text{th}}$ - 114$^{\text{th}}$), including only 5 instances of each target word.

E Measuring Sentiment on the Backbenches

To construct that sentiment estimate for the House of Commons, we take the following steps:

1. for the policy area of interest, designate the seed word (e.g. 'nhs' for the "Health")

2. embed that seed using ALC as described above (specifically using `GloVe` embeddings and the original Khodak et al. (2018) **A** matrix). We now have an embedding for every instance of the term. Aggregate those embeddings to party-rank-month (so if Tory backbenchers mention 'eu' twice in July 2015, we take the average embedding of those two mentions)

3. using inner product as a measure of similarity, compare that aggregate embedding to the embedding of words in a sentiment dictionary—i.e. the embeddings of words thought to connote positive or negative valence. The specific dictionary we use for this purpose is the Affective Norms for English Words (Warriner, Kuperman and Brysbaert, 2013), preprocessed and operationalized in the way described by Osnabrügge, Hobolt and Rodon (2021, ftn 9).

4. for a given (averaged) embedding for the party-rank-triple, calculate its valence as its mean similarity to the set of positive terms *plus* minus one multiplied by the mean similarity to the set of negative terms.

5. finally, we rescale those valences within party and term of interest (i.e. Tory-backbench and Tory-cabinet sentiment towards a given term over the time series is scaled 0-1, and the same is done for Labour-backbench and Labour-cabinet sentiment).

This general approach is inspired by the word embeddings association test (WEAT) of Caliskan, Bryson and Narayanan (2017) for measuring bias in text. Their approach uses cosine similarity as the measure of similarity rather than the inner product. Although widely used, this approach has been criticized for depending too heavily on the relative frequency of the seed word and the target words (Ethayarajh, Duvenaud and Hirst, 2019; van Loon et al., 2022). This dependence may arise in part due to the standardization by magnitude in the cosine measure. We might expect this problem to be less severe in our setting because we are comparing against two different embeddings of the same word, which—in this example—are used frequently by both groups.

At the time of writing, this is an active area of study and so out of an abundance of caution, we ran the study using both the conventional cosine similarity and the inner product (which does not standardize). We then presented the inner product results in the main paper as it has less well-defined patterns. Figure E.5 presents the results using cosine similarity. Researchers looking to use a similar design should consult the latest literature on the topic for guidance.

F Experiments with decision variables

While our approach does not require any active tuning of parameters, there are nevertheless choices to be made. To better guide practice, we ran a number of experiments. First, we briefly revisit our `Trump` vs. `trump` example from Section 4. Recall that our task is to classify

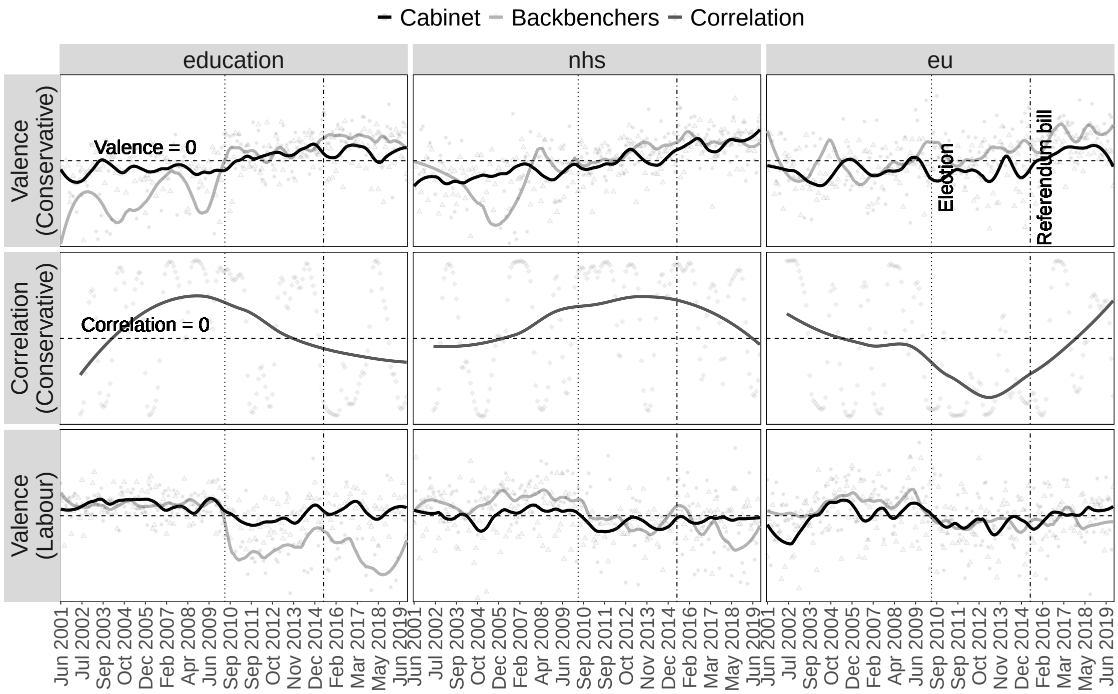


Figure E.5: The equivalent of Figure 9 using cosine similarity.

ALC embeddings of individual mentions of the term `Trump/trump` in our NYTs corpus, into one of the two senses of the term—a *supervised task*. We use Stanford GloVe—window size of 6, 300 dimensions—as our pre-trained embeddings along with its corresponding A matrix as estimated by (Khodak et al., 2018). We further use k-means clustering to assign each single-instance ALC embedding into one of $k = 2$ clusters. To evaluate performance we use three common clustering metrics:

- **Homogeneity**: maximized ($h_i = 1$) when each cluster contains only members of a single class; minimized ($h_i = 0$) when each cluster contains a random assortment of members.

- **Completeness**: maximized ($c_i = 1$) when all members of a given class are assigned to the same cluster; minimized ($c_i = 0$) when members are randomly assigned to clusters.

- **V-measure**: a weighted combination –harmonic mean– of homogeneity and completeness; the more homogeneous and complete a given clustering, the higher this score –

bounded between 0 (worst) and 1 (perfect).

And focus on two modeling choices:

1. **Context window size**: this refers to the window size of the contexts used to estimate the ALC embeddings which we set to 2, 6 or 12.
2. **Stopword removal**: we evaluate the effect of removing stopwords at the point of preprocessing the contexts used in estimating the ALC embeddings.[14]

Figure F.6 plots these results. First, we observe that irrespective of the size of the window or removal of stopwords, ALC is capable, with varying degrees of success, of distinguishing between the two senses. However, we do observe that for this particular task, larger windows and the removal of stopwords can be helpful, showing marginally higher scores across all three metrics.

[14]We use `quanteda`'s list of stopwords.

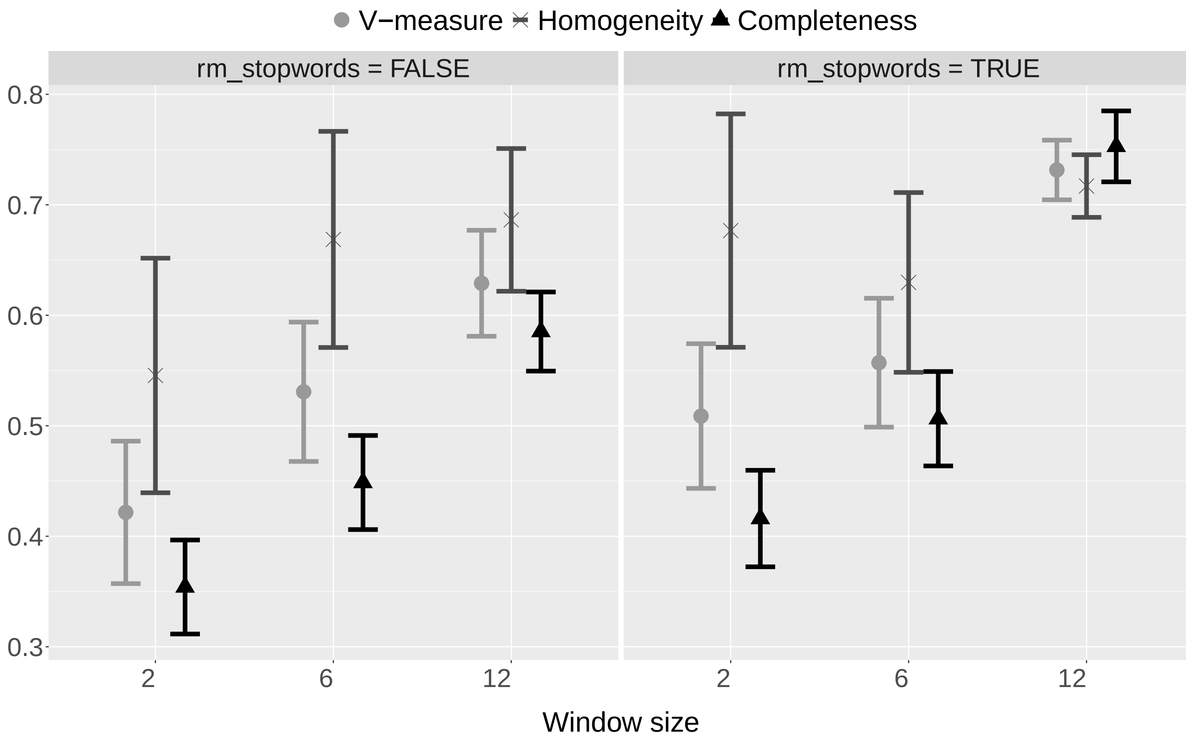


Figure F.6: Performance on a supervised classification task as a function of context window size and leaving/removing stopwords.

We next turn to our Congressional Records corpus–Session 107 - 114.[15] We evaluate performance along three metrics:

- **Reconstruction**: similarity between the estimated ALC embeddings of a set of terms and their corresponding embedding in the set of pre-trained embeddings. The higher this similarity, the better ALC "reconstructs" the 'true' underlying embeddings (i.e. low bias).

- **Nearest neighbors**: overlap in nearest neighbors to the estimated ALC embeddings, as measured by the Jaccard Index, across the various model specifications. The higher this overlap, the lower the variance as a function of model specification.

[15]We use a larger swathe of the Congressional Records than in our main example in the paper in order to train high quality locally-fit embeddings.

- **Substantive**: temporal trends in partisan differences for a set of political terms. We define the partisan difference for a given term during a given session of Congress as the cosine similarity between the two party –Republican and Democrat– ALC embeddings. For each term we then have a time series spanning the eight sessions of Congress that constitute our corpus. The goal is to evaluate how these trends compare across model specifications. We quantify this using Pearson correlation. The higher the Pearson correlation, the lower the variance –in the substantive interpretation of results– as a function of model specification. We call these "substantive" in that they capture the types of relationships researchers are often interested in e.g. temporal variation in group differences.

We narrow our comparisons to a set of 'political' terms: `democracy`, `freedom`, `equality` `justice`, `immigration`, `abortion`, `welfare`, `taxes`, `republican`, `democrat`–as in Rodriguez and Spirling (2022). And focus on the same two modeling choices as in the previous experiment, with some differences in implementation:

1. **Context window size**: we train a separate "full" locally-fit `GloVe` embeddings model for each of three window size 2, 6 and 12 and estimate the corresponding **A** matrices. All models include any feature that appears at least 10 times in the corpus and are at least 3 characters long. Models are trained for 25 iterations.[16] We also evaluate Stanford GloVe (window of size 6, 300 dimensions) embeddings—downloaded from online—but in this case estimate a "locally-fit" **A** matrix i.e. using our Congressional Records corpus as input for the estimation rather than using Khodak et al. (2018)'s A.

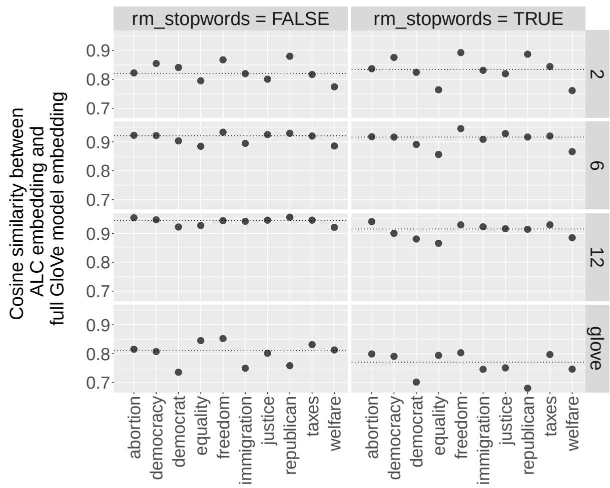
2. **Stopword removal**: as with the `trump`/`Trump` experiment, we evaluate the effect of removing stopwords at the point of preprocessing the contexts used in estimating the ALC embeddings. We leave stopwords in when estimating the full embeddings models as is standard practice.

[16] We use the R package `text2vec` to estimate "local" GloVe embeddings.

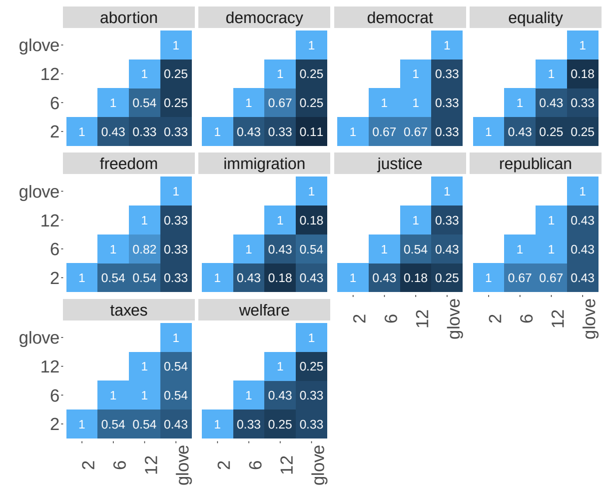
Our results are summarized in Figure F.7. In Figure F.7a we observe that ALC performs its intended task well—in most cases extremely well—namely to reconstruct the original embeddings. Across all "local" model specifications we see an average cosine similarity across the ten terms of above 0.8 with larger window models (6 and 12) achieving an average above 0.9. Results suggest avoiding smaller windows (< 5) could be advantageous, although not strictly necessary. Even using `GloVe` pre-trained embeddings with a localized $\mathbf{A}$ matrix achieves excellent results. In other words, lacking enough data to train their own embeddings models, users can reasonably resort to pre-trained embeddings trained on (large and broad) corpora. Removing stopwords does not improve results, indeed if anything results are generally slightly worse. This is not altogether surprising in that stopwords were not removed when estimating the full embeddings model nor, more importantly, when estimating the $\mathbf{A}$ matrix. The latter takes care of reweighting dimensions in a way that mitigates the prevalence of stopwords.

In Figure F.7c we observe the overlap in the top 10 nearest neighbors across the various model specifications as measured by the Jaccard Index. Results suggest significant overlap across models, even between Stanford `GloVe` pre-trained embeddings and local models.

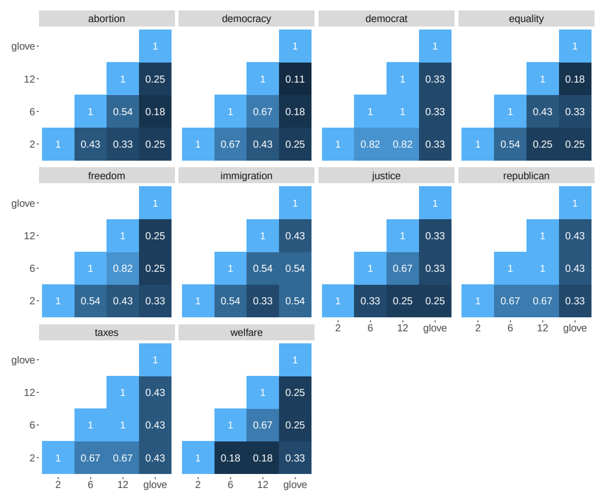
Figure F.7d plots the trends in partisan differences over time. The appropriate way to read this plot is to compare for each term the time trends across model specifications – so compare plots within a column. We observe significant overlap across specifications for all terms. We can further quantify this using Pearson correlation (see Figure F.7f). Again, we observe very high (above 0.9) correlations across all models, including between Stanford `GloVe` and the set of locally trained models.



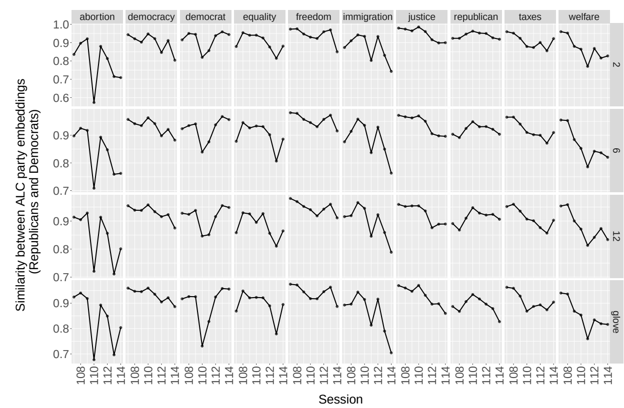
(a) Reconstruction



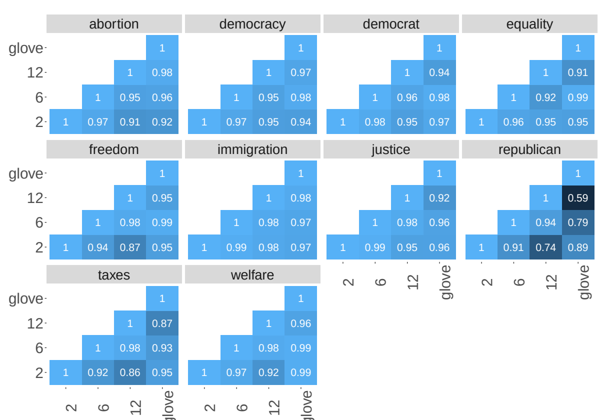
(b) Nearest neighbors (Jaccard Index)



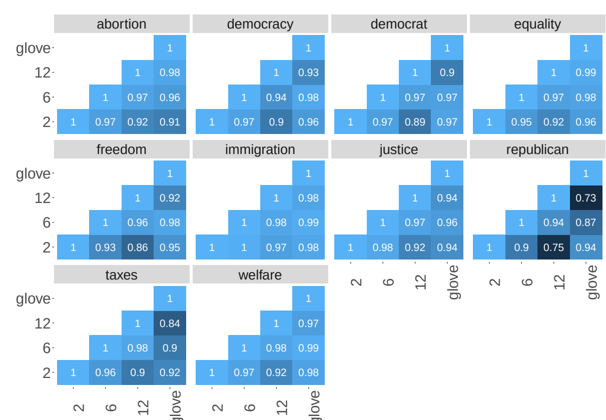
(c) Nearest neighbors (Jaccard Index) w/o stopwords



(d) Substantive (visualization)



(e) Substantive (correlation)



(f) Substantive (correlation) w/o stopwords

Figure F.7: Performance as a function of context window size and leaving/removing stopwords.

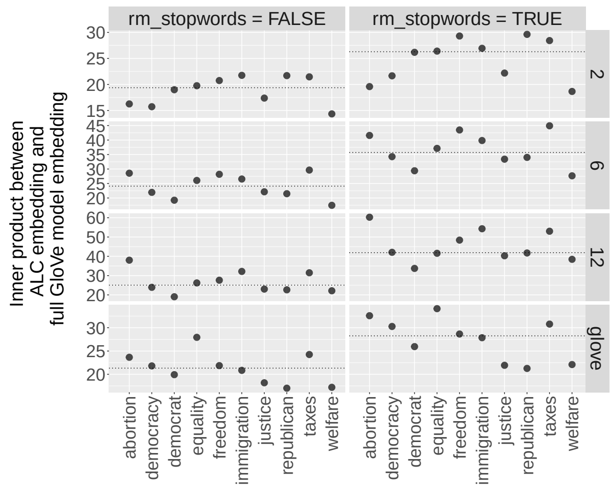
In addition to our experiments above, we also looked at the following decision variables:

- Similarity metric: cosine similarity versus inner product. The former only cares about the angle between vectors while the latter cares about the angle and magnitude. Indeed, cosine similarity is equivalent to normalizing the inner product by the magnitude of the vectors. It is uncommon to use inner product as a similarity metric due in part to its sensitivity to document length.

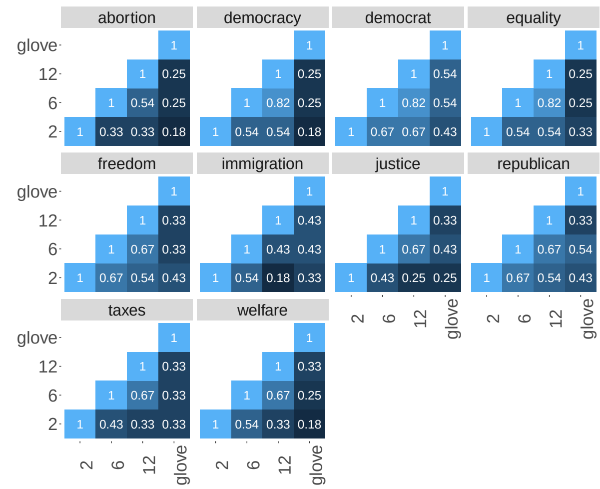
- Stemming: it's often the case that nearest neighbors show various terms with the same stem e.g. "enforcing" and "enforce". A user can easily group these by averaging the cosine similarities to nearest neighbors with the same stem.[17]

Figure F.8 summarizes results using the inner product as the similarity metric. Results are generally more variable then when using cosine similarity but we nevertheless still observe significant overlap in nearest neighbors and high correlations in temporal trends across model specifications. Unless there's a very problem-specific reason to use inner product we suggest users follow common practice and use cosine similarity as a metric.

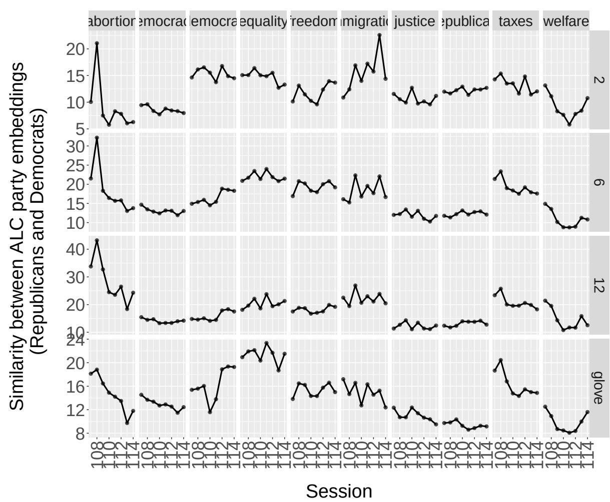
[17] A practical matter to note here is that averaging across terms with the same stem may introduce noise through low-frequency misspelled terms with low cosine similarities to the target word. To avoid this we suggest subsetting candidate terms to correctly spelled words –this can be automated– when using stemming.



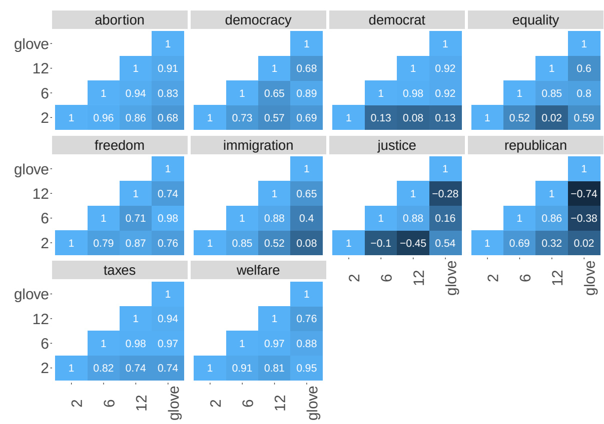
(a) Reconstruction



(b) Nearest neighbors (Jaccard Index)



(c) Substantive (visualization)



(d) Substantive (correlation)

Figure F.8: Performance as a function of similarity metric and stemming.

Figure F.8 summarizes results using stemming – in this case it only makes sense to look at our nearest neighbors metric. With some exceptions we observe significant overlap across models. While stemming may help group similar terms users should keep in mind it often comes at the cost of interpretability.

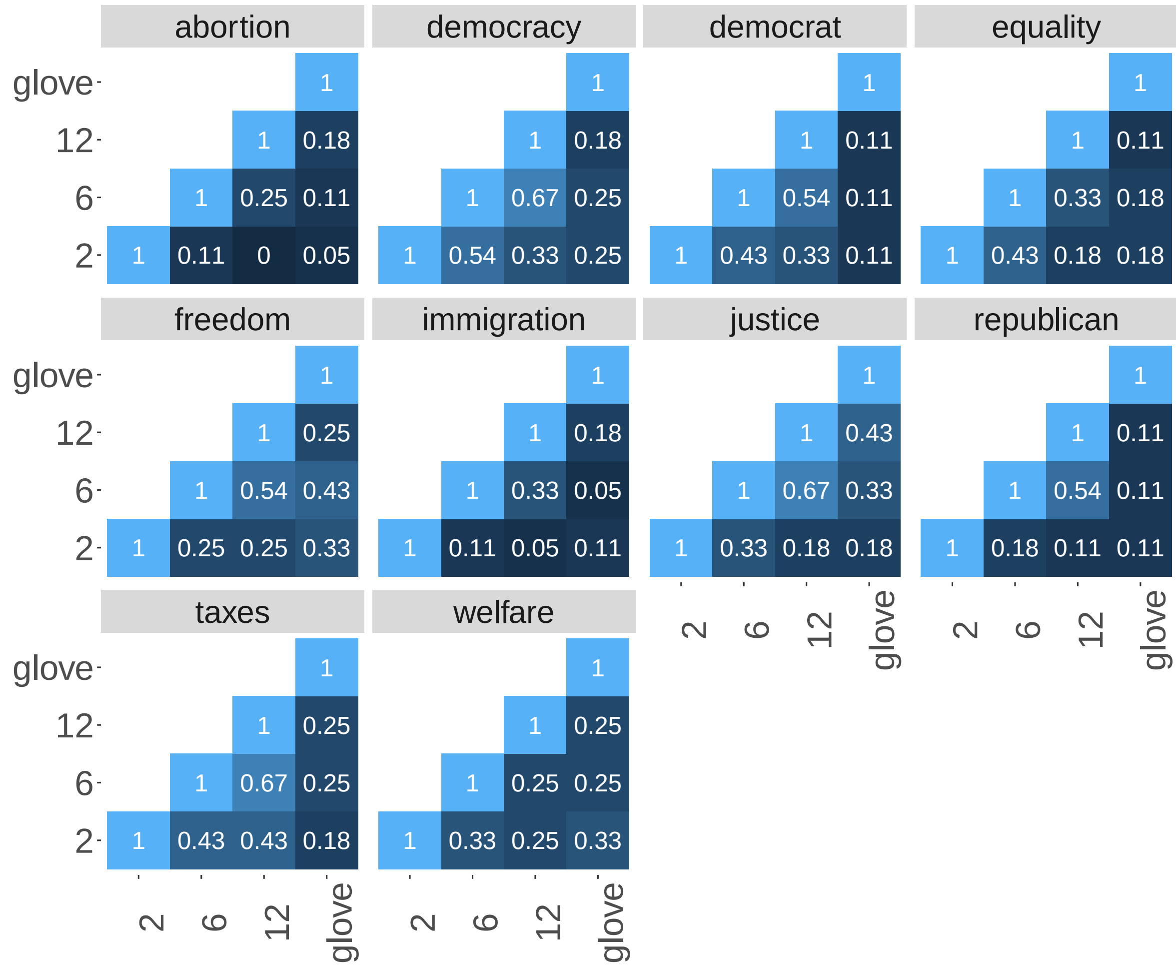
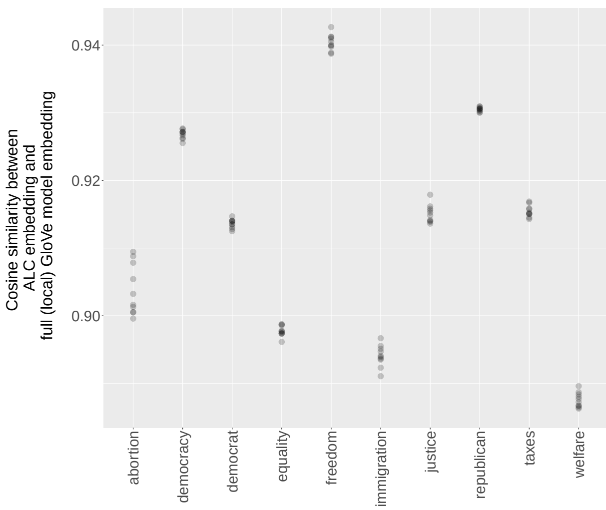


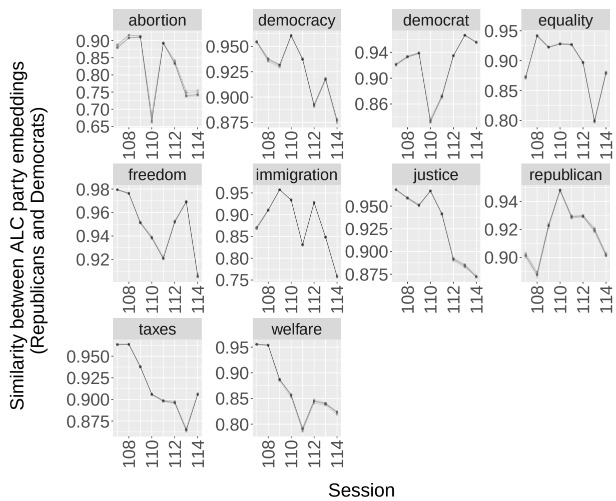
Figure F.9: Substantive (correlation)

G A Matrix uncertainty

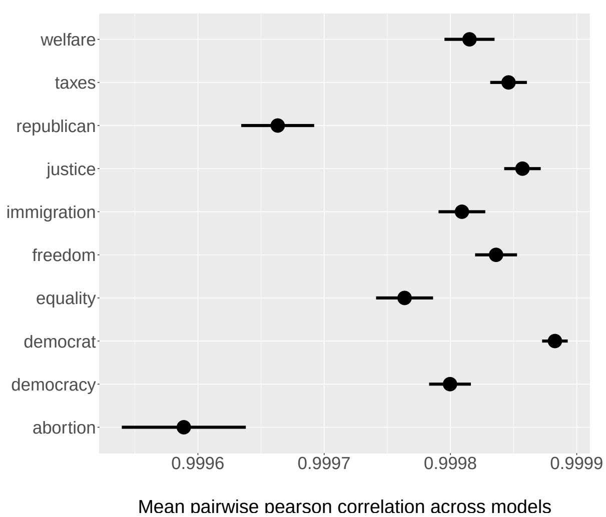
Throughout the examples in our paper we have assumed the transformation matrix **A** to be known and fixed, ignoring any uncertainty that may arise as result of having to estimate **A**. In the experiment that follows we evaluate how reasonable said assumption may be. We again use the Congressional Records (Sessions 107–114) as our corpus and a locally estimated GloVe model with context window size of 6. We next estimate 10 different **A** matrices for 10 bootstrapped samples of the corpus and apply the same evaluation framework as described in Section F. While our objective in Section F was to compare results across various model specifications, in this case it is to compare results across the 10 estimates of **A**. Across all metrics–*reconstruction*, *nearest neighbors* and *substantive*– we see remarkably indeed negligible differences (see Figure G.10). Users should consider uncertainty in the calculation of the **A** matrix as a second-order concern, and unlikely to be consequential for topline results.



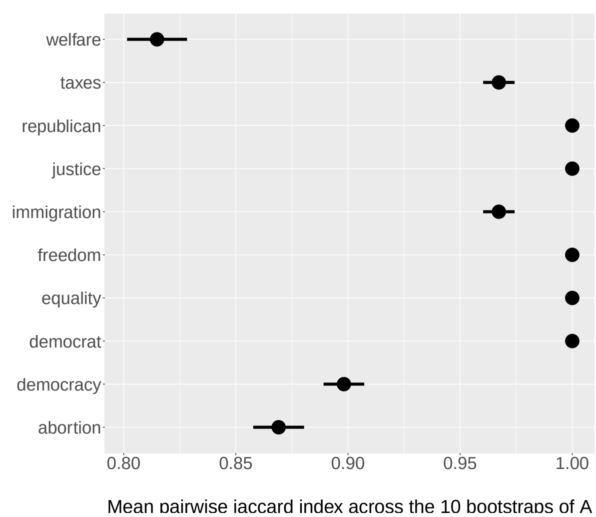
(a) Reconstruction



(b) Substantive (visualization)



(c) Substantive (correlation)



(d) Nearest neighbors (Jaccard Index)

Figure G.10: Performance as a function of A matrix estimation.

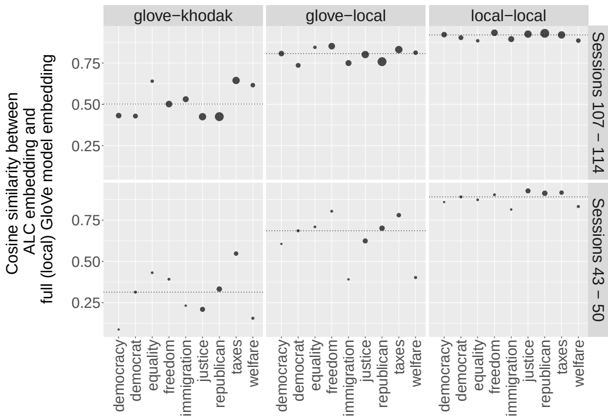
H Variation over time

One concern users of the proposed approach may have is that pre-trained embeddings estimated on modern texts are ill-suited to study older texts. Take for example our replication of Rodman (2020) above where we use Stanford `GloVe` embeddings to study a corpus that includes data spanning 161 years. To shed light on this concern we look at two subsets of the Congressional Records: Sessions 43–50 (1873–1889) and Sessions 107-114 (2001–2017). For each subset we estimate a local embeddings model—context window size 6—and corresponding **A** matrix. We want to evaluate how well results using Stanford `GloVe` pre-trained embeddings match results using these local models with the expectation that we should observe larger differences when applying Stanford `GloVe` pre-trained embeddings to study the earlier sessions of Congress. To round off our comparison we evaluate differences using Stanford `GloVe` pre-trained embeddings with Khodak et al. (2018)'s **A** matrix and the same embeddings but with a locally estimated **A** matrix. So, for each period we have the following combinations of models:

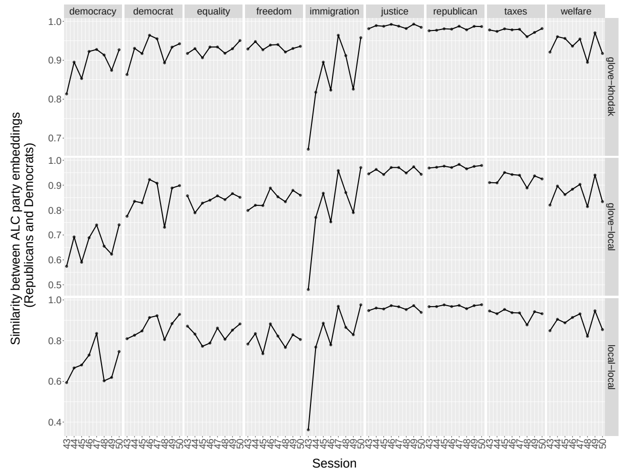
- **local - local**: locally trained embeddings on the corresponding corpus and a locally estimated A matrix.
- **GloVe - local**: Stanford GloVe pre-trained embeddings and a locally estimated A matrix.
- **GloVe - GloVe**: Stanford GloVe pre-trained embeddings and the corresponding A matrix estimated by Khodak et al. (2018).

We again use the evaluation framework laid out in Section F. Figure H.11 summarizes our results. A couple of general patterns emerge. As expected, locally trained models with their corresponding **A** matrices show the best performance –irrespective of time period— across all metrics— reconstruction, nearest neighbors and substantive. Nevertheless, Stanford `GloVe` pre-trained embeddings with a locally estimated **A** matrix show remarkably strong performance in terms of our reconstruction metric and nearest neighbors even for the earlier period—albeit somewhat worse than when employed to analyze more recent texts. As expected, Stanford `GloVe` pre-trained embeddings with Khodak et al. (2018)'s $\mathbf{A}$ shows somewhat lower performance in terms of reconstruction and nearest neighbors. However, turning to our substantive metric, with some exceptions correlations are high across all models, suggesting that even Stanford `GloVe` with Khodak et al. (2018)'s $\mathbf{A}$ performs well in capturing the general trends in the underlying data. What should readers make of these results? Given a large corpus, locally trained embeddings and corresponding A matrix is desirable. However, for smaller corpora, using large pre-trained embeddings models such as Stanford `GloVe` embeddings, will be more than adequate in most cases, ideally with a locally trained $\mathbf{A}$ matrix—given enough data—but not necessarily so.[18]

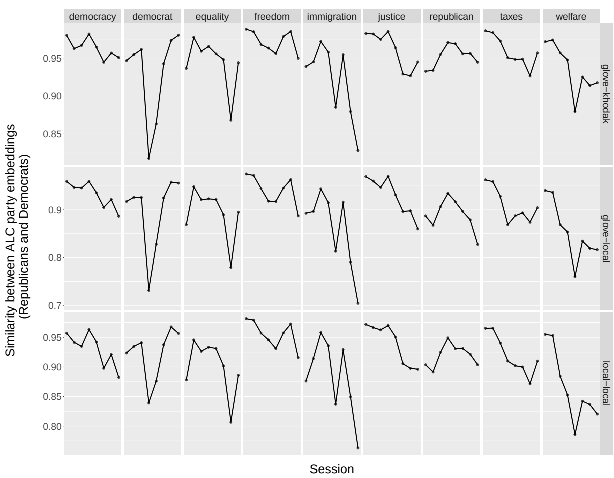
[18]Note, data requirements to train an $\mathbf{A}$ matrix are orders of magnitude lower than to train a full embeddings model. For a D dimensional embedding space, the former requires estimating $D \times D$ coefficients, whereas the latter requires estimating $V \times D$, with V representing vocabulary size and generally $V >> D$.



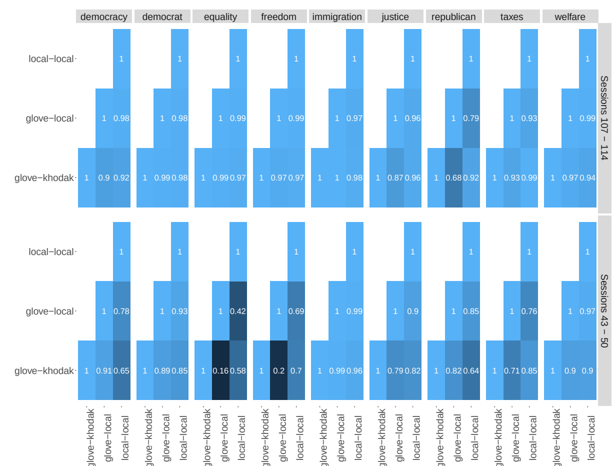
(a) Reconstruction



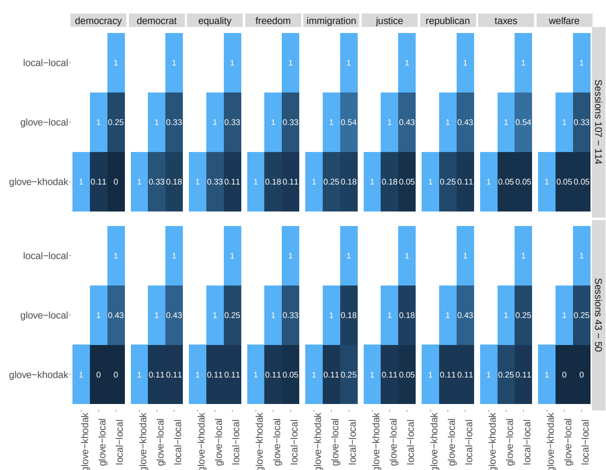
(b) Substantive (visualization)



(c) Substantive (correlation)



(d) Nearest neighbors (Jaccard Index)



(e) Nearest neighbors (Jaccard Index)

Figure H.11: Performance as a function of time.

I Software

To facilitate applying the methods presented in this paper we put together an R package – `conText`. The main function `conText` follows generic `R` `lm( )` and `glm( )` syntax in terms of $\sim$ operator. Please refer to the quick start guide to get started using the package. As with any package, we had to make a couple of design decisions that are worth noting here. First, ALC embeddings are computed using the available pre-trained context word embeddings. If a given context word is not available in the provide pre-trained embeddings, then that context word is simply ignored and the average is taken over the set of available context embeddings. Second, we've found that in practice limiting the candidate nearest neighbors to the set of words in the provided contexts, significantly reduces noise (non-sensical nearest neighbors such as misspelled words etc.). Whenever exploring nearest neighbors you can use the parameter `candidates` to delimit the set of nearest neighbors. Finally, we have made available—or simply more accessible—the GloVe pre-trained embeddings used in most of the examples in this paper along with their corresponding transformation matrix.[19] We are often asked when is it appropriate to use these pre-trained embeddings and their corresponding transformation matrix rather than estimate ones own. Unfortunately, there is no hard-and-fast rule for this, it comes down to how distinct you think your corpus is relative to the corpus used to train these embeddings (Wikipedia 2014 and Gigaword 5).

References

Caliskan, Aylin, Joanna J Bryson and Arvind Narayanan. 2017. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356(6334):183–186.

Ethayarajh, Kawin, David Duvenaud and Graeme Hirst. 2019. "Understanding undesirable word embedding associations." *arXiv preprint arXiv:1908.06361* .

[19]The original GloVe embeddings computed by the Stanford NLP Group can be found here while the original transformation matrix computed by Khodak et al. (2018) can be found here.

Khodak, Mikhail, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon M. Stewart and Sanjeev Arora. 2018. A la carte embedding: Cheap but effective induction of semantic feature vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne, Australia: Association for Computational Linguistics pp. 12–22.

Osnabrügge, Moritz, Sara B Hobolt and Toni Rodon. 2021. "Playing to the gallery: Emotive rhetoric in parliaments." *American Political Science Review* 115(3):885–899.

Rodman, Emma. 2020. "A timely intervention: Tracking the changing meanings of political concepts with word vectors." *Political Analysis* 28(1):87–111.

Rodriguez, Pedro L. and Arthur Spirling. 2022. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *The Journal of Politics* 84(1):101–115.

van Loon, Austin, Salvatore Giorgi, Robb Willer and Johannes Eichstaedt. 2022. "Regional Negative Bias in Word Embeddings Predicts Racial Animus–but only via Name Frequency." *arXiv preprint arXiv:2201.08451* .

Warriner, Amy Beth, Victor Kuperman and Marc Brysbaert. 2013. "Norms of valence, arousal, and dominance for 13,915 English lemmas." *Behavior research methods* 45(4):1191–1207.